



University
of Glasgow

Alexander, Marc Gabriel

(2011) *Meaning construction in popular science: an investigation into cognitive, digital, and empirical approaches to discourse reification.*

PhD thesis.

<http://theses.gla.ac.uk/3171/>

Copyright and moral rights for this thesis are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Meaning Construction in Popular Science

An Investigation into Cognitive, Digital, and
Empirical Approaches to Discourse Reification

Marc Gabriel Alexander

Submitted for the degree of Doctor of Philosophy

*English Language, College of Arts
University of Glasgow*

November 2011

ABSTRACT

This thesis uses cognitive linguistics and digital humanities techniques to analyse abstract conceptualization in a corpus of popular science texts. Combining techniques from Conceptual Integration Theory, corpus linguistics, data-mining, cognitive pragmatics and computational linguistics, it presents a unified approach to understanding cross-domain mappings in this area, and through case studies of key extracts, describes how concept integration in these texts operates.

In more detail, Part I of the thesis describes and implements a comprehensive procedure for semantically analysing large bodies of text using the recently-completed database of the *Historical Thesaurus of English*. Using log-likelihood statistical measures and semantic annotation techniques on a 600,000 word corpus of abstract popular science, this part establishes both the existence and the extent of significant analogical content in the corpus. Part II then identifies samples which are particularly high in analogical content from the corpus, and proposes an adaptation of empirical and corpus methods to support and enhance conceptual integration (sometimes called conceptual blending) analyses, informed by Part I's methodologies for the study of analogy on a wider scale. Finally, the thesis closes with a detailed analysis, using this methodology, of examples taken from the example corpus. This analysis illustrates those conclusions which can be drawn from such work, completing the methodological chain of reasoning from wide-scale corpora to narrow-focus semantics, and providing data about the nature of highly-abstract popular science as a genre.

The thesis' original contribution to knowledge is therefore twofold; while contributing to the understanding of the reification of abstractions in discourse, it also focuses on methodological enhancements to existing tools and approaches, aiming to contribute to the established tradition of both analytic and procedural work advancing the digital humanities in the area of language and discourse.

ACKNOWLEDGEMENTS

Above all else, the existence of this thesis is owed to the grace, knowledge, keen perception, genuine good humour and unfeigned enthusiasm of both Catherine Emmott and Christian Kay. I am delighted to have had the fortune to have been taught by both of them prior to the start of this thesis, and I am more grateful than I can say to have had the honour of working with each of them over the past six years. Both deserved a better student but made the best of what they had, and in so doing have radically transformed the way I think. It is to them I dedicate this thesis, with simple thanks.

In writing the thesis, I have benefited from a great deal of friendship, encouragement and enthusiasm from a wide range of people, and I have been affected profoundly by the support and good humour of Robyn Bray, Fraser Dallachy, and Andrew Struan. This last in particular has provided genuine support and ceaseless chatter for the past eleven years, and deserves more sincere acknowledgement than this sentence provides.

Jean Anderson, who had the benevolence to employ me for almost three years (and the geniality to make it pass in the blink of an eye), introduced me to new people, new methodologies and new ways of thinking about the humanities in the best manner possible; by pointing me towards data and encouraging me towards results from the ground up. I hope she can recognize the debts this thesis owes to her as continuations on that same theme.

I am indebted to my examiners, Lisa Lena Opas-Hänninen and Wendy Anderson, for the invaluable advice they have given me, their enthusiasm for the thesis' content, and the attention to detail they have shown. However, all errors and infelicities that remain in the thesis are, of course, entirely my own.

I have, or have had, the great pleasure of working with the *Historical Thesaurus of English* project, the *Enroller* project, the *Parliamentary Discourse* project, the *STELLA* project, the *Stylistics, Text Analysis and Cognitive Science* project, and the *Language and Sentiments of Their Times* project, all of which meant I was fortunate to gain multiple insights into a wide range of disciplines and methodologies. Collaborators, friends, and coworkers in various projects have been a delight to work with (including some already mentioned above), and include Jean Anderson, Ellen Bramwell, Flora Edmonds, Catherine Emmott, Dorian Grieve, Johanna Green, Carole A. Hough, Christian Kay, Hannah Little, Alison Phipps, Andrew Prescott, Anthony J. Sanford, Sulman Sarwar, Fiona Stewart, Andrew Struan, and Catherine Wild.

Three further supervisors, John Corbett, Michael Burke, and Michael K. C. MacMahon have all provided ungrudging help and good-natured support at various points since 2006. The staff at Glasgow's Department of English Language (now within the School of Critical Studies), my Honours semantics, pragmatics, digital humanities, and discourse analysis students, and friends and colleagues from PALA, ALLC/ADHO and PhilSoc have all provided intellectual stimulation, unrestrained support, and easy friendship for a number of years, and have shown genuine interest in what I have had to say even when I wasn't sure how best to say it. Beyond those already mentioned above, particular thanks go to Anthony Johnson, Ilkka Juuso, Jukka Tyrkkö, Ville Marttila, Carla Suhr, Mary Ellen Ryder, Dawn Archer, Paul Rayson, Jane Roberts, Kathryn Allan, Philip Durkin, Bettelou Los, and Jonathan Hope from outside the University of Glasgow, and Jeremy J. Smith, Jennifer Smith, Katie Lowe, Jenny Bann, Elizabeth Robertson, Irené Wotherspoon, Ann Gow, and Stephen Woodruff from various departments of that institution. Pauline Maridor, Alison Bennett, Helen Gilday, Gill Carr, Kathleen Johnson, Richard Codd, and Elaine Wilson have all been paramount examples of kind, knowledgeable, and friendly university administrators, and my immense good fortune in working with them has been the source of much jealousy from friends at other institutions.

The Carnegie Trust for the Universities of Scotland funded this thesis, and deserve special thanks for their generous, cheerful and commonsensical assistance every single time I dealt with them. The projects above which I have been part of were funded variously by the Trust, the Arts and Humanities Research Council, the Joint Information Systems Committee and the Higher Education Academy, who have all contributed in various ways to some of the ideas proposed here.

Finally, my mother, Mary Kerr, did not live to see this thesis completed, but would have been inordinately proud of it nonetheless (and regardless of its content). My sister, Susan Turetta, has also been endlessly supportive, enthusiastic and keen to see its completion. My thanks go to both of them, for everything they have done for me.

TABLE OF CONTENTS

Chapter 1: Introduction.....	12
1.i: Research Questions.....	14
1.ii: Reification, Metaphor and Analogy.....	15
1.iii: Popular Science and Popularizations.....	18
1.iii.1: A Definition of Popular Science.....	19
1.iii.1.a: ‘Interpretation’.....	19
1.iii.1.b: ‘General Audience’.....	20
1.iv: Related Work.....	20
1.v: This Thesis.....	22
Part I: Macro-Level.....	24
Chapter 2: Empirical Outline.....	25
2.i: Distant Reading.....	25
2.i.1: Introduction.....	25
2.i.2: Cognitive Linguistics and Computational Techniques.....	27
2.i.3: Similar Work.....	32
2.ii: Methodology.....	34
2.ii.1: Introduction.....	34
2.ii.2: The Data.....	35
2.ii.2.a: <i>The Corpora</i>	35
2.ii.2.b: <i>The Historical Thesaurus of English</i>	39
2.ii.2.c: <i>Acronymic Summary</i>	42
2.ii.3: The Process.....	43
2.ii.4: Some Technical Detail.....	48
2.ii.5: Polysemy.....	49
2.ii.6: Rationale and Alternatives.....	51
2.iii: Methodological Overview.....	54
Chapter 3: Key Analogical Domains.....	56
3.i: Introduction.....	56
3.ii: Analysis: Core Texts.....	57
3.ii.1: The Fabric of the Cosmos.....	57
3.ii.1.a: <i>FC Key Domains</i>	57

3.ii.1.b: FC: Analogical Categories.....	64
3.ii.1.c: FC: Concordances.....	68
3.ii.1.d: FC: Summary	79
3.ii.2: PW: Parallel Worlds.....	79
3.ii.2.a: PW: Key Domains	79
3.ii.2.b: PW: Analogical Categories.....	80
3.ii.2.c: PW: Concordances	82
3.ii.2.d: PW: Summary	85
3.ii.3: MP: The Music of the Primes.....	85
3.ii.3.a: MP: Key Domains.....	85
3.ii.3.b: MP: Analogical Categories.....	85
3.ii.3.c: MP: Concordances	87
3.ii.3.d: MP: Summary.....	91
3.ii.4: WP: Warped Passages.....	92
3.ii.4.a: WP: Key Domains	92
3.ii.4.c: WP: Summary	97
3.iii: Comparison Texts.....	98
3.iii.1: C-NT.....	98
3.iii.2: C-AC.....	99
3.iii.3: C-GEB.....	100
3.iv: Conclusion.....	101
Part II: Discourse-Level.....	105
Chapter 4: A Cognitive Approach to Reification.....	106
4.i: Introduction	106
4.ii: ‘Key’ Extracts.....	107
4.ii.1: Outline	107
4.ii.2: Global Analogical Content: Key Domain Extracts.....	107
4.ii.2.a: Textual Clusters of Analogical Content: FC	109
4.ii.2.b: Textual Clusters of Analogical Content: MP	113
4.ii.3: Local Analogical Content	113
4.ii.3.a: A Sample of Local Analogical Content: PW.....	116
4.ii.3.b: A Sample of Local Analogical Content: WP.....	118
4.ii.4: Summary.....	118
4.iii: Quantitative to Qualitative.....	120

4.iv: Conceptual Integration Theory	I21
4.iv.1: Literature.....	I23
4.iv.2: The Process of Integration.....	I24
4.iv.2.a: <i>The Nature of the Blend</i>	I25
4.iv.3: Mental Spaces	I27
4.iv.4: Application.....	I28
4.iv.5: The Creation of a Blend	I29
4.iv.5.a: <i>Selection</i>	I29
4.iv.5.b: <i>Compression</i>	I30
4.iv.5.c: <i>Mappings</i>	I32
4.iv.6: Governing Principles	I33
4.iv.6.a: <i>Overarching Goals</i>	I33
4.iv.6.b: <i>Governing Principles for Compression</i>	I34
4.iv.6.c: <i>Other Governing Principles</i>	I35
4.iv.7: Representational Considerations.....	I36
4.iv.8: Criticisms.....	I37
4.iv.8.a: <i>A Post Hoc Theory</i>	I38
4.iv.8.b: <i>A Broad Theory</i>	I40
4.iv.8.c: <i>Granularity</i>	I41
4.iv.8.d: <i>Use</i>	I42
4.v: Reification Analysis in Practice	I42
4.v.1: The Identification of Input Spaces.....	I43
4.v.1.a: <i>Relevance Theory and Input Spaces</i>	I45
4.v.1.b: <i>Relevance Theory Principles for Input Spaces</i>	I48
4.v.2: The Constitution of Input Spaces.....	I50
4.v.3: The Internal Makeup of Input Spaces	I53
4.v.4: An Analytic Example.....	I54
4.v.4.a: <i>Extract Selection</i>	I54
4.v.4.b: <i>Identifying Input Spaces</i>	I55
4.v.4.c: <i>Naming and Populating Input Spaces</i>	I57
4.v.4.d: <i>Space Relationships</i>	I59
4.v.4.e: <i>Concept Integration</i>	I62
4.vi: Conclusion	I85
Chapter 5: The Discourse of Reification.....	I86
5.i: Introduction.....	I86

5.ii: Blends and Blending.....	187
5.ii.1: Blending the Key Abstracts.....	188
5.ii.1.a: <i>Initial Methodological Notes</i>	188
5.ii.1.b: <i>Analysis</i>	192
5.ii.2: Detailed Notes.....	197
5.ii.2.a: <i>Warped Passages</i>	197
5.ii.2.b: <i>The Fabric of the Cosmos</i>	199
5.ii.2.c: <i>The Music of the Primes</i>	199
5.ii.3: Multimodality	199
5.ii.3.a: <i>The Lakovian Principles</i>	200
5.ii.3.b: <i>Prior Usage in Conceptual Integration</i>	201
5.ii.3.c: <i>Conclusion</i>	202
5.iii: Reification in the Key Extracts	203
5.iii.1: Outline.....	203
5.iii.2: Reification	203
5.iii.3: Types of Blended Space	205
5.iii.4: Textual Progression Markers	209
5.iv: Conclusion	210
Chapter 6: Conclusion	211
6.i: Summary	211
6.i.1: Chapter 2: From Semantics to Corpora	212
6.i.2: Chapter 3: From Corpora to Analogical Domains	212
6.i.3: Chapter 4: From Analogical Domains to Conceptual Blends	212
6.i.4: Chapter 5: From Conceptual Blends to Discourse.....	213
6.i.5: Overall: From Semantics to Discourse.....	213
6.ii: Effectiveness and Shortcomings.....	215
6.iii: Future Work.....	216
6.iv: Envoi.....	217
Bibliography.....	218
Appendices.....	233
Appendix I: Python Programs and SQL Queries.....	234
WordList.py.....	235
WordList.py: Annotated	236
MakeString.py.....	238

MakeString.py: Annotated.....	238
OrderedList.py.....	239
OrderedList.py: Annotated.....	240
LineSplit.py.....	241
LineSplit.py: Annotated.....	242
CFDCorpusSearch.py.....	243
CFDCorpusSearch.py: Annotated.....	243
SQL: LogLikelihood Datasheet.....	244
SQL: LogLikelihood Datasheet: Annotated.....	245
SQL: Occurrences of Top 100 Domains.....	247
SQL: Occurrences of Top 100 Domains: Annotated.....	249
Appendix II.1: Vital Relations in Conceptual Integration Theory.....	251
Appendix II.2: Parallel Worlds.....	256
<i>a: Full Text.....</i>	<i>256</i>
<i>b: Annotated Text.....</i>	<i>259</i>
<i>c: Input Mental Spaces.....</i>	<i>261</i>
<i>d: Semantic Input Space Groups.....</i>	<i>262</i>
<i>e: Grouped Mental Spaces.....</i>	<i>264</i>
<i>f: Mental Spaces and Elements.....</i>	<i>265</i>
<i>g: Relationship Diagram.....</i>	<i>267</i>
<i>h: Blended Spaces.....</i>	<i>269</i>
<i>i: Blending Commentary.....</i>	<i>271</i>
Appendix II.3: Warped Passages.....	273
<i>a: Full Text.....</i>	<i>273</i>
<i>b: Annotated Text.....</i>	<i>276</i>
<i>c: Grouped Mental Spaces.....</i>	<i>280</i>
<i>d: Mental Spaces and Elements.....</i>	<i>280</i>
<i>e: Relationship Diagram.....</i>	<i>282</i>
<i>f: Blended Spaces.....</i>	<i>284</i>
<i>g: Blending Commentary.....</i>	<i>284</i>
Appendix II.4: The Fabric of the Cosmos.....	286
<i>a: Full Text.....</i>	<i>286</i>
<i>b: Annotated Text.....</i>	<i>289</i>
<i>c: Grouped Mental Spaces.....</i>	<i>291</i>
<i>d: Mental Spaces and Elements.....</i>	<i>292</i>

<i>e: Relationship Diagram.....</i>	<i>293</i>
<i>f: Blended Spaces.....</i>	<i>295</i>
<i>g: Blending Commentary.....</i>	<i>295</i>
Appendix II.5: The Music of the Primes.....	297
<i>a: Full Text.....</i>	<i>297</i>
<i>b: Annotated Text.....</i>	<i>30I</i>
<i>c: Grouped Mental Spaces.....</i>	<i>305</i>
<i>d: Mental Spaces and Elements.....</i>	<i>306</i>
<i>e: Relationship Diagram.....</i>	<i>308</i>
<i>f: Blended Spaces.....</i>	<i>3I0</i>
<i>g: Blending Commentary.....</i>	<i>3II</i>
Appendix III (on attached CD)	

Much scientific progress – whether it involves theorizing about invisible forces, studying the behaviors of atoms, or trying to characterize the nature of private mental experience – depends on generating new ways of describing and conceptualizing phenomena that are not perceivable through the senses.

Boroditsky and Ramscar 2002:185

CHAPTER I: INTRODUCTION

We have suggested that [...] we tend to structure the less concrete and inherently vaguer concepts (like those for the emotions) in terms of more concrete concepts, which are more clearly delineated in our experience.

Lakoff and Johnson 1980:112

[...] as soon as one gets away from concrete physical experience and starts talking about abstractions or emotions, metaphorical understanding is the norm.

Lakoff 1993:205

Following the 1980 publication of George Lakoff and Mark Johnson's groundbreaking *Metaphors We Live By*, a foundational text in cognitive linguistics, it has been near-axiomatic in that field that human beings, as embodied minds perceiving and construing the mental, social and physical worlds around them, understand abstractions in terms of concrete entities. While this is a well-explicated concept,¹ no stylistic studies have yet aimed to establish both the extent and operation of such abstract-as-concrete construals in a corpus of discourse. It is this lacuna that the present thesis, situated at the intersection of cognitive linguistics and the digital humanities, sets out to address, by describing this linguistic strategy of cross-domain concrete-abstract semantic mappings in the popularization of science. In so doing, and in the tradition of digital language studies, it makes a wide range of suggestions and improvements to empirical methodologies and analytical techniques in these two fields, by taking both a digital-empirical and cognitivist approach to the nature of construal within a particular subgenre of popular science texts, and in so doing demonstrates both the existence of the reification-mapping strategy in

¹ See, for example, the book-length exposition of the centrality of embodiment in Lakoff and Johnson 1999 ('the very structure of reason itself comes from the details of our embodiment', *ibid*:4), and psychological studies such as Boroditsky and Ramscar 2002, Feldman 2006, Biggs 1990 and 2006, and Coulson 2008.

that genre and the multiplex nature of the stylistic and linguistic construal of complex ideationally-driven discourse.

The rationale behind the focus of the present study is detailed below, but, as a brief overview, popular science texts focusing on highly-abstract concepts (a mathematical exposition of number theory, a physics text on parallel universes, and two books on string and quantum theory) were selected as representative texts of discourses which are focused on explaining abstract concepts to a non-specialist audience, and therefore provide the greatest potential for the analysis of non-literary analogy. As analogy is not exclusively a two-domain process (that is, analogies are not axiomatically singular entities, but rather tend to co-occur so that more than one analogy can explain the same concept), a theory which involves the analysis of multiple conceptual sources to form coherent conceptual outputs was chosen within the cognitive linguistics framework, namely Conceptual Integration Theory (also known as conceptual blending). As a corpus linguist as well as a cognitivist, I also took as a cornerstone of the present study the necessity to ground any analysis wherever possible in the tradition of corpus linguistics, prioritising the empirical strengthening of any analysis of discourse, and in so doing ‘take advantage of the computer’s ability to test hypotheses over large stretches of text’ (Sinclair 2004:16). As a result, the present study is a cognitive analysis of analogy with a significant focus on methodology and heuristics, taking the opportunity to enhance this cognitive analysis wherever possible with new and innovative technologies and datasets. In the title of this thesis, I deliberately use the formulation ‘approaches’, and in so doing hope to highlight that my contribution to the field aims to be in the area of such *approaches* as well as in a more traditional discourse analysis. In so doing, I reaffirm that even in the cognitive arena, which treats introspection as

a valid methodology and source of data,² as much attention as possible should be placed on John Sinclair's claim that research should be:

founded on principles of maximal attainable objectivity, procedures that are as replicable as possible. While I think we should always be guided ultimately by our own personal taste and intuition, the more of our reasoning that we can place in the public domain, the better. 'Ferocious empiricism' was Patrick Hanks' phrase for my attitude to research [...] and I am proud to wear that tag.

Sinclair 2007:6

At those stages where more space is placed on methodological improvements in the present thesis than on the eventual analysis, this is considered a strength rather than a flaw, both as a combination of the rightful commitment to this 'ferocious empiricism' wherever possible, and also as a recognition that 'those working in the digital humanities have long held the view that application is as important as theory [...] typified] by critical and interpretative tools and analytic algorithms' (Schreibman, Siemens and Unsworth 2004:xxv). While a detailed discourse analysis of a particular corpus is undertaken in this thesis, and a number of linguistic conclusions reached, these form a test case for the methods of analysis, given that the pathways to those conclusions are, at times, more significant than the destination.

I.1: RESEARCH QUESTIONS

In accordance with the overview above, the overall research framework of this thesis is to ask in what ways digital humanities techniques can add to the cognitive linguistic analytic process of analogy. In order to narrow and focus

² 'The new insights into the system of conceptual structuring in language that have been coming from the relatively recent tradition of cognitive linguistics have rested mainly on the methodologies already standard in the field of linguistics overall: introspection in conjunction with theoretical analysis.' Talmy (2007:xii). It is worth noting that Talmy argues against this perspective, although not with any great strength (compare Gibbs 2006a), instead concluding that other methods of data collection, such as corpus data, can have some equal standing with introspection. This is not the position taken in this thesis.

that approach for the purposes of research, the main research question addressed concerns the general workings of the process of making an abstract concept ‘concrete’ within highly abstract popular science. To give this a clear analytic statement, through which wider methodological issues can be approached, the research question asked is: *How does cross-domain mapping within popular science texts operate as a strategy of reifying discourse construal?* This question gives rise to three hypotheses which are examined and proven throughout the course of the work. These are:

Hypothesis 1: Popular science texts present and construe abstractions by means of construing them with reference to multiple and non-congruent concrete domains.

Hypothesis 2: The cognitive description of this cross-domain mapping can be undertaken using a cross-disciplinary model which can deal with multiple separate domains being construed within the same stretch of text.

Hypothesis 3: The cognitive analysis of such texts can be methodologically grounded by digital humanities resources and techniques, such as corpora and semantic datasets.

Hypothesis 1 is the subject of Chapters 2 and 3. Hypothesis 2 is the subject of Chapters 4 and 5. Hypothesis 3 is examined throughout the thesis, and is particularly addressed through the theoretical discussion in Chapter 2 and the methodological discussions in Chapter 4.

I.II: REIFICATION, METAPHOR AND ANALOGY

This thesis uses two particular words, *reification* and *analogy*, as superordinate terms which aim to avoid some issues of interpretation with alternative terms.

The first of these is *reification*, which is here used, in its straightforward sense, to mean ‘make an abstract thing concrete’. While *concretization* is the term most often found in cognitive linguistics, where it is discussed at all (eg in Lakoff and Johnson 1980), *reification* is here preferred in order to distance the present thesis from the presupposed conclusion that the construal of the abstract necessarily involves that which is ‘concrete’. *Reification*, literally meaning *making into a thing*, has the same meaning in the main, but avoids the explicit presupposition somewhat. Its corollaries of *concrete* and *abstract* in the sense of parts of a conceptual framework, and semantic classes, are minimally defined here respectively as ‘that which is perceived through physical senses’, and ‘that which is distant from those perceptions’.

Rather more unusually, this thesis uses the term *analogy* to refer to the construal of one domain in terms of another, or the phenomenon of ‘seeing’ one thing in terms of another. This is untypical because, for many theorists, the term *metaphor* would be the natural way to refer to this process, and whole swathes of cognitive linguistics centres around this term and its associated theories. It is, in fact, because of this multiplicity of theories³ that *metaphor* as a term is avoided where possible. The range of metaphor research has meant that there is some agreement on what constitutes a prototypical metaphor (such as ARGUMENT IS WAR), but some difficulty when one reaches marginal cases (dead metaphors, primary metaphors, structural metaphors, orientational metaphors, ontological metaphors, correlation metaphors, resemblance metaphors, and so on). When considering the whole field of figurative language, there is a lack of agreement on whether metaphor and metaphorical processes such as metonymy cover the whole field, or if different phenomena should be posited

³ In addition to what can be called the Lakovian approach, characterized by George Lakoff and his followers and collaborators (Lakoff 1987, 1993, Lakoff and Johnson 1980, 1999, 2003, Lakoff and Turner 1989, and other in the same vein, such as Kövecses 2002), there is a key Anglo-European approach which differs from the Lakovian approach in a number of ways (Steen 2002, 2007, Semino 2008), and then some other alternative approaches, such as those of Gentner (see Gentner and Bowdle 2008 for an overview) or Sperber and Wilson (2008, *inter alia*).

to account for various different aspects of the field. In the approach which this thesis takes, Conceptual Integration Theory, conceptual metaphor is a particular sub-type of its basic unit, the conceptual blend. This has been acknowledged by a recent joint statement by George Lakoff, the pioneer of conceptual metaphor theory, and Gilles Fauconnier, the pioneer of conceptual blending, which states ‘the research programs developed for metaphor and blending [are] mutually reinforcing and often deeply intertwined’ (2010).

It follows that if metaphor can be seen as a hyponym of a blend, or at the least an intertwined sister concept, it is useful to delineate the specific concerns of this present thesis by generally avoiding the term *metaphor* itself for the particular phenomenon of reificatory analogy. More empirical and comparative work is needed in this field to systematically delineate modes of figurative language depending on the conceptual structure they instantiate, on the subtle and complex cline which ranges between metonymy, metaphoricity, allegory, and beyond. Some steps have been taken here by Line Brandt, in her 2010 thesis (318ff), which attempts in part to schematize blending into a semiotic framework based on purpose, although this is unfortunately not a wholly linguistic theory or one which can be systematically related to linguistic structure in a way which this thesis would require. Instead, as a compromise, *analogy* is used to refer to the phenomenon this work addresses; the expository use of a ‘seeing-as’ relationship, as Croft and Cruse (2004:209) characterize conceptual blends, which operates at a discourse level across multiple clauses and across yet larger structural units, such as paragraphs or even chapters.

Note, though, that in no way does the present thesis deny the existence of metaphor or the long tradition of its analysis in cognitive linguistics. Where appropriate, particular instances of metaphoricity with broad agreement on the term given to it, such as diachronically-shifting dead metaphors, are referred to as such. The concern here is rather with avoiding the danger of engaging with all of metaphor, rather than the particular phenomenon of reificatory

analogy. It is hoped that readers interested in metaphor will nonetheless find this thesis of relevance and of some interest.

I.III: POPULAR SCIENCE AND POPULARIZATIONS

As a study of discourse semantics, a suitable corpus of text is required on which to base the analyses in this thesis. From the perspective of Hypothesis I in §I.i above, it can be expected that analogical discourse which presents abstract ideas to a reader would be relatively common; anything addressed to a general audience in a non-physical domain (society, economics, mathematics, emotions, faith, sound, and so forth) would be expected to use concrete analogies to aid comprehension. In order to address the status of this hypothesis, it is reasonable to use a text genre which is heavily concerned with abstract concepts and in their exposition and comprehension. As a result, it was decided to take the texts for analysis from pedagogical texts aimed at adults, as these are both focused on comprehension and also sufficiently advanced to be likely to discuss a range of abstract concepts. Not using pedagogical texts for younger readers also had the advantage of avoiding issues around reading age and educational curricula.

To narrow text selection even further from general adult non-narrative non-fiction, the largest⁴ representative genre of that field which has a significant focus on abstract concepts is that of popular science; a genre which, in a crude definition to be revised in the following section, aims to interpret specialist scientific knowledge to a general audience. The thesis therefore focuses on this genre, and the corpus for analysis will be drawn from texts here with a high probability of including highly-abstract concepts, by choosing texts which focus on aspects of science which are particularly non-concrete. This excludes the life sciences, concerned with physical biology, and the chemical sciences, concerned with elements and their compounds, as both of these are distinctly concrete in

⁴ As detailed by the UK Publishers' Association, personal correspondence.

focus. Instead, the major branches of science which are most concerned with abstractions and theoretical models are mathematics and physics,⁵ which is the specific sub-genre which this thesis will focus on.

Below, the field of popular science is outlined and a definition of it as a genre is given, while the actual mechanics of selecting a precise corpus of this genre is left for Chapter 2, which gives the methodological overview for Part I of the thesis.

I.III.I: A DEFINITION OF POPULAR SCIENCE

The initial definition given above of popular science was a genre which aims to interpret specialist scientific knowledge to a general audience. Two parts of this definition need further exposition below, namely the key phrases *interpretation* and *general audience*.

I.iii.I.a: ‘Interpretation’

The term *interpretation* highlights the fact that any scientific knowledge, before being communicated to the public, has already been deduced from analytic results and also published in scientific proceedings such as journals. Therefore, the process of popularization involves taking that which is to be popularized (the ‘learned text’), and re-expressing and re-interpreting it. The idea of facts being interpreted by a science writer is by no means a new one, and science itself is often presented as an interpretation of evidence; this last epistemological view is known as ‘constructivism’ (Knorr-Cetina and Mulkey 1983). Wheatley (1991:10) gives two main principles of constructivism:

Principle one states that knowledge is not passively received, but is actively built up by the cognizing subject.... Principle two states that the function of cognition is adaptive and serves the organization of the experiential world, not the

⁵ Although these are presented as separate fields here, theoretical physics and mathematics shade into one another, to the extent that the distinction between them in some sub-fields is difficult to make – for example, Stephen Hawking, one of the best-known living physicists, was for thirty years before his retirement Lucasian Chair of Mathematics at Cambridge.

discovery of ontological reality.... Thus we do not *find* truth but construct viable explanations of our experiences.

The view above that interpretation is an essential part of popularization is therefore consistent both with major theories of the epistemology of science as well as in linguistics.

I.iii.I.b: 'General Audience'

Briefly, the term *general audience* in the definition above can be interpreted in many different ways. *General audience* is a privative term; it exists in opposition to a number of other audiences which are implicitly non-general. These audiences can be non-expert (a popularization for people interested in a field but not highly knowledgeable about it) or non-specialist (for people knowledgeable about the general field and cognate areas but not in the particular topic of the popularization). It is therefore preferable to replace the privative 'general' with instead the relative term *outsider* in the definition, to give a revised version of: 'a genre which aims to interpret specialist scientific knowledge for "outsiders" to the scientific field under discussion'.

I.IV: RELATED WORK

Section §2.i.3 discusses prior work in the digital methodology area of Part I of the thesis, while §4.iv does the same for Conceptual Integration Theory in Part II. This section, therefore, is correspondingly brief and covers other studies of popular science within stylistics.

Major linguistic studies of popular science have generally focused, in a systemic-functional and applied-linguistic sense, either on the comparison between popularizations and 'insider' texts (eg Corbett 1992 and 2006, or Myers 1990a, 1990b, 1994) or on the discussion of popular science as a genre (from the science communication point of view, eg Neiman 2000, or from a critical discourse perspective, such as van Dijck 1998). These have also,

overwhelmingly in the case of the studies cited, been primarily focused on the life sciences, which this thesis explicitly avoids (see further §1.iii above). Some fruitful studies of science fiction (particularly Stockwell 1992 and 2000) have made occasional side references to popular science, but have not engaged with it as a genre. A further point to be made is that much existing work in the field has focused on popular science in a comparative mould; the source-text-target-text model where theorists focused on language as a social semiotic rather than a cognitive artifact have examined the textual changes between a source text in the scientific literature and a target text in a popularized field. These studies have mainly been within applied linguistics with an eventual pedagogical aim (eg Corbett 1990 and Myers 1990a). Corpus stylistic studies of popular science are similarly rare outside of EFL studies, and the only one of significance (that is, the only one which makes more than a passing reference) is Butler's 1990 study of modal verbs in popular science. I have found no further in-depth linguistic studies of analogy in the semantic construction of popular science, and particularly not in the cognitive field. Reasons for this lacuna can be attributed to the relative lack of interest in scientific non-narratives in stylistics, when compared with the natural focus many scholars place on either fiction or political texts.⁶ There are also fewer texts available in comparison to these much larger genres, and almost no texts available in the public domain for large-scale corpus work, due to the relative youth of the genre. A combination of these factors alongside the recent nature of the analytic procedures has resulted in this gap in the literature.

However, genre issues aside, there have been other studies which take similar approaches to texts in the style of this thesis, and readers are referred to the more comprehensive notes on this found in the later sections outlined above.

⁶ This also reflects the university structure of many scholars in the field, who are often part of general English or English Literature departments, and so focus on fiction and literary stylistics.

I.V: THIS THESIS

The thesis is divided into two parts: Part I, a ‘top-down’ macro-level approach to this research, taking a broad-based corpus linguistic methodology to a large corpus; and Part II, a ‘bottom-up’ discourse-level approach, which uses cognitive linguistic methodologies to progress from lexicology to semantics and thus to discourse. This present chapter, the introduction, and Chapter 6, the conclusion, surround these two parts. However, although the parts are treated as separate textually, they connect in a number of ways, and in the methodological approaches used in each inform the other. The two parts are therefore not functionally separable.

Beyond the main text, there are three sets of appendices included in the thesis, one for each part (Appendix I and Appendix II respectively) and a final electronic appendix (Appendix III), containing the large datasets used in the analysis of Part I. (For copyright reasons, the electronic appendix unfortunately cannot include full copies of the original texts, which are recent and in-print texts from commercial publishers.) Appendix III is provided on CD-ROM with this thesis, and is also available online at <http://phd.seldomend.com>, which also makes available the programs designed in Part I and printed in parts of Appendix I.

Each section and sub-section of the thesis is headed for the purposes of cross-reference. The traditional format of hierarchical heading numbering is not wholly appropriate for this thesis, as a later section will introduce category numbers from the *Historical Thesaurus* dataset. These category numbers are of the form 01.04.09.07.03 (this one being the precise reference for the colour *red* in that dataset). In order to avoid confusion between such category numbers and cross-references in this thesis, hierarchical sections of the present work are instead referred to in accordance with the following schema:

Number.RomanSmallcase.Number.Letter.Number.Letter.Number

Giving, for example, a heading 3.ii.1.c.3.a.3, which identifies the subsection below which deals with a concordance analysis of the lexeme *line* in one particular corpus text. The advantage of this mixed system is not just that of distinguishing headings from *Thesaurus* categories, but also provides a useful mixed identifier; the first element is the chapter number, the second the chapter sub-division, and then anything following the major chapter sub-divisions is alternately numbered and lettered. Finally, as used above, note that all cross-references are prefixed by the section symbol, §.

PART I: MACRO-LEVEL

Words, words. They're all we have to go on.

Stoppard 1967:1.348-349

CHAPTER 2: EMPIRICAL OUTLINE

This chapter introduces and explicates the corpus-driven technique of semantic profiling developed for this thesis, and applies it to a corpus of popular science in order to describe the extent of analogy in the corpus texts. The chapter outlines the data used, the process of analysis, and the rationale for the methodological decisions made, following a detailed introduction and survey.

2.1: DISTANT READING

2.1.1: INTRODUCTION

This chapter and the following are digital humanities investigations of a corpus of highly-abstract popular science texts. In order to address the question of the operation of analogical reification in such texts, and also to describe ways of getting results from these corpora which can be used in more detailed cognitive analysis, the corpus needs to be analysed from both a semantic and from a wide-scale perspective. In effect, Part I of this thesis is an exercise in *distant reading*, Franco Moretti's (2000) term for the opposite of literary study's close reading, where:

[...] distance is, however, not an obstacle, but a specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models.

Moretti (2005:2)

Unlike Moretti's ideal grand qualitative perspective on hundreds of books, this thesis will instead focus on a relatively small corpus, which offers a balance between one large enough for macro-level distant reading, and one small enough to offer a manageable set of data for the later discourse-level studies undertaken in Part II.

The reason any sort of macro-level investigation is required, as stated above, is that it is currently hypothetical that popular science of the type identified is actually analogical on a broad scale. It seems in an intuitive sense to be true, but this has not been established by any systematic independent analysis based on linguistic form. These two chapters therefore aim to empirically establish the range and nature of semantic domains in use in a corpus of popular science texts. They do so through an original series of Python programs running on plain-text corpora, and by applying a semantic profiling technique using a database derived from the 2009 Glasgow *Historical Thesaurus of English*. By so doing, it empirically identifies and establishes the extent of analogy within a corpus of book-length texts and provides a procedure for doing the same.

The core hypothesis to be addressed in these two chapters is that popular science texts systematically employ multiple semantic domains which are strictly of no relevance to the scientific data under discussion in the text, for the purpose of analogy. This has a consequential hypothesis that a concordance-based analysis of these *prima facie* ‘irrelevant’ domains will provide evidence that such domains are actually analogically employed to concretize the discussion of the scientific content of the text. Also, by comparison of the core corpus with a secondary one, the final hypothesis to be discussed is that popular science operates differently in this area from two similar text types, namely a scientific text aimed at a learned community and popularized humanities research.

Beyond these research questions, the other main aim of this section is to provide an empirical and data-driven foundation from which to undertake the discourse-level domain-based stylistic and semantic analyses in the second half of this thesis. These follow on from the initial quantitative analyses of textually-global systematic analogical domains in §3.i below to qualitative stylistic analyses based on shorter extracts of text but informed by the methodology established in this chapter.

Accordingly, the remainder of this section outlines the relationship between the cognitive approach to meaning (in both the semantic and discursive sense) and computational approaches to text analysis, and briefly surveys methodological work close to that undertaken in this thesis.

2.1.2: COGNITIVE LINGUISTICS AND COMPUTATIONAL TECHNIQUES

Until recently, the cognitive approach has not often been seen as one which simultaneously engages with computational approaches to language analysis. However, there has been some work undertaken in recent years by a group of scholars who aim to provide links between the two, and either to use digital techniques to provide more robust underpinnings to the field, or to employ computational resources to enhance (and in some ways to ground) cognitive analyses of text.

This growth has been due to some of the prior criticisms of cognitive linguistics. While the field is a strong and active one, it inherited certain theoretical and methodological biases from the mentalist and generative traditions it emerged from and as a reaction to. As outlined in Chapter 1 above, such methodologies have been criticized as relying too much on introspection and the intuitions of native speakers at the expense of empirical data (see the detailed discussions by Gibbs (2007:2ff), Talmy (2007:xiff), Evans and Green (2006:78off), and Steen (2007:19ff), amongst others). As Gibbs says (2007:4):

The fact that we think we can introspect about the inner workings of our minds does not mean that such intuitions, even if trained, are either consistent or accurate. [...] people can significantly vary from one day to the next in reporting their beliefs or knowledge, even for simple things like the names of all the birds or furniture they know.

This lack of rigour has resulted in results which, though ‘intuitively appealing’ (Evans and Green 2006:780), were easily criticized ‘for lacking a clear set of methodological decision principles’ (*ibid*). Gibbs’ position is well-

founded when he says that although ‘introspections can be valuable sources for constructing hypotheses, we must always be cautious in accepting any individual analyst’s linguistic judgments’ (2007:3). In particular, cognitivist analyses are often open to the charge of being post-hoc in their orientation; a feature of interest is identified by an analyst using instinct and only afterwards given a cognitive explanation, rather than the theory leading to predictions which can be validated or discarded in order to feed into improved analytical methods.

These criticisms are valid in some ways and questionable in others. Notably, the distinction between the post-hoc nature of cognitivist approaches to language and the equally post-hoc nature of other fields such as CDA, systemic-functionalism, and many others, is outwith the scope of this thesis (see, for example, Widdowson 2004:109). There is an argument to be made that certain linguistic phenomena require interpretation by a human mind, which automatically renders them necessarily post-hoc; some linguists (for example in the field of conceptual blending; see further Chapter 4) take no issue with the post-hoc nature of their analyses, and others make no claim to be part of those scientific fields where Popperian notions of falsifiability hold sway (see Everett 2002 and, with regards to stylistics, Short *et al* 1998 and Jeffries 2000).

This is a somewhat intractable argument which is based around a particular individual’s theoretical orientation towards the study of language and the nature and scope of linguistics; it will not be pursued further here. Instead, the rather more relevant point to make is that, regardless of its *a priori* status, cognitive linguistics, and cognitive stylistics, can benefit from increasing its empirical footing in general, following the perspectives of corpus linguistics and the digital humanities. This outlook not only addresses certain of the concerns above with regards to the subjectivity of analysis, but also gives analysts powerful further apparatuses to use in their work. This is not to say that current analyses are invalid or inappropriate, but rather that in the case of

the evolution of any approach to language, new tools become available to supplement existing ones as they are developed. This broad-scale, tool-based approach to language and discourse has been a core principle of stylistics since before cognitive linguistics emerged as a major field of linguistic analysis; for example, Leech and Short in 1981 stated:

The more a critic wishes to substantiate what he says about style, the more he will need to point to the linguistic evidence of texts; and linguistic evidence, to be firm, must be couched in terms of numerical frequency [...] If challenged, I ought to be, and can be, in a position to support my claim with quantitative evidence.

(1981:46-7)

This approach is as valid to cognitive stylistics as it is to general stylistics, and to cognitive linguistic analyses in general. Evidence-based approaches can assist theoretical, speculative and emerging perspectives on the nature of language as a discursal product of the human mind; the only possible problem arising from this principle occurs if there exist no appropriate empirical models to be employed. This is in accordance with what is said by Gries *et al* 2005:

It is by now generally acknowledged in Cognitive Linguistics that, in accordance with its cognitive commitment, hypotheses and constructs should be backed up by converging evidence from multiple sources (for pertinent programmatic statements cf., e.g., Lakoff and Johnson 1999:79–81, Langacker 1999b:26) as well as firmly anchored to the empirically most adequate kinds of evidence.

Taking this at face value, and in conjunction with the arguments above, it therefore follows that, should there be no appropriate empirical methodology for undertaking a particular investigation, an analyst should try where possible to develop one.

The major problem which arises here is the essential divide between corpus/empirical approaches to language and the cognitive approach. The core principles of the latter are well-established enough to allow scholars to identify

the common threads running throughout cognitive linguistic work. As Geeraerts and Cuyckens state in the recent *Oxford Handbook of Cognitive Linguistics*:

[...] three fundamental characteristics of Cognitive Linguistics can be derived: the primacy of semantics in linguistic analysis, the encyclopedic nature of linguistic meaning, and the perspectival nature of linguistic meaning. The first characteristic merely states that the basic function of language involves meaning; the other two characteristics specify the nature of the semantic phenomena in question. The primacy of semantics in linguistic analysis follows in a straightforward fashion from the cognitive perspective itself: if the primary function of language is categorization, then meaning must be the primary linguistic phenomenon.

(2007a:5)

Statements to this effect can be found in all of the major overviews of the field, including Croft and Cruse's statement that 'linguistic knowledge – knowledge of meaning and form – is basically conceptual structure' (2004:2) and the following from Geeraerts:

The foundational point is simply that language is all about meaning. As it says in the Editorial Statement of the very first issue of the journal *Cognitive Linguistics*, published in 1990, this approach sees language 'as an instrument for organizing, processing, and conveying information' – as something primarily semantic.

(2006:3)

The issue here is that, by contrast, corpus linguistics does not directly deal with meaning. As a computational approach to language, it can only deal with that on which computers can operate. In other words, it is concerned with *strings*, sequences of characters which programs can count, divide into words, and so forth (see, for example, Sinclair 2004). Computational linguistics has built substantially on these foundational restrictions, and programs can now recognize words and sentences appropriately, tag them with part of speech

information to a reasonable tolerance (although never entirely accurately), and perform other summative and presentational work to the extent that examinations of lexis and syntax are incomparably enhanced by the use of corpus methods (see, *inter alia*, Sinclair 2004, Biber *et al* 1999, and, in a pioneering early example, Quirk *et al* 1985).

One issue which remains with the approach, though, is a lack of engagement with semantics. In particular, meaning is not predictable as a lexical or grammatical phenomenon; in fact, in the cognitive view, lexis and grammar are rather surface representations of meaning. This means that cognitive linguistics, in common with other stylistic, semantic and meaning-focused approaches, has to use lexis as a proxy for meaning; to discover facts about (say) the representation of the body in discourse using corpora requires either selecting the most common lexis to express parts of the body (a difficult task, but possible), attempting to search for all the possible lexical realizations of the body, or working on a combination of non-corpus and corpus sources. In this way, the use of a proxy often aims for a corpus to be searched *representatively*; that is, using a series of terms which are considered to be appropriately representative of the semantic domain under consideration (choosing key terms, either from the analyst's own knowledge or through using other data sources such as thesauri or electronic dictionaries; see Garretson and O'Connor 2007). This is entirely appropriate pragmatically, however it has meant that many prior discourse corpus studies have in fact been lexical studies aiming at semantic/discoursal representativeness rather than being purely semantic. Previous work combining corpus and cognitive linguistics is almost exclusively lexical or syntactic (for example, the full title of a recent book on the topic is *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis* (Gries and Stefanowitsch 2006), while the important article *A Case for a Cognitive Corpus Linguistics* (Grondelaers *et al* 2007) has important theoretical implications but demonstrates its case by examining the distribution of a single Dutch particle). This observation is not intended as criticism; there were

previously no frequently-used alternatives to primarily non-semantic approaches to corpora, although this thesis aims to make a contribution towards expanding the possibilities of meaning-based corpus work by using materials from lexical semantics as proxies.

The present thesis has been undertaken within a particular research context, and at a particular time, where there are substantial opportunities for exploiting linguistic resources in combination. These opportunities include the maturity of natural language processing tools and their open-source availability, and the completion of the 44-year-long *Historical Thesaurus* project (see §2.ii.2.2 below). This chapter, and the following, therefore aim to demonstrate a methodology of computationally engaging with semantic domains using such resources, and, in addition, to contribute to answering the core research question of abstract meaning reification.

2.1.3: SIMILAR WORK

Similar work in this area is limited, having been hampered in the past by the lack of a sufficiently fine-grained and comprehensive semantic resource for English (or any other language). The only comparable work is in semantic tagging, which is not what this thesis aims to do with the *Historical Thesaurus*, but nonetheless the general principles of this field are broadly analogous to it.

The major system currently used for semantic tagging, and the only one to be found in corpus linguistics of English, is the Lancaster USAS (see Wilson 1993 and Archer *et al* 2002; USAS stands for the UCREL Semantic Annotation System, while UCREL currently stands for the University Centre for Computer Corpus Research on Language).⁷ The USAS system consists of 235 categories, based on the *Longman Lexicon of Contemporary English* (McArthur 1981), and uses a machine-training system in combination with an existing pre-tagged lexicon of 37,000 words. It operates thus:

⁷ The acronym, well-known in the field, has stayed the same while the project title has changed. There is no longer a direct link between the acronym and the title.

The tagged text is fed into the main semantic analysis program, which assigns semantic tags representing the general sense field of words from a lexicon of single words and an idiom list of multi-word combinations (e.g. *as a rule*). The tags for each entry in the lexicon and idiom list are arranged in general rank frequency order for the language.

(Wilson 1993)

As USAS is a tagger (unlike the identification system used in this thesis) it has a multiple set of disambiguation methods to then contextually reduce the semantic possibilities of each word down to a single tag, which is appended in turn to each lexeme in the text. The USAS system is commercially successful and widely used within its field, but it is unfortunately not appropriate for the style of analysis to be undertaken in this thesis, which requires identification of semantic domains to a conceptually-relevant degree of granularity.⁸ The USAS tagset has 235 categories, one thousandth that of the *HT*, and thus has broad-scale categories, such as *N2 Mathematics*, *WI The universe*, and so on, which correspondingly contain the full range of all the concepts which could be placed under those headings. These are reflections of the aims of McArthur (to provide a learner's lexicon organized by meanings) and USAS' intentionally broad-scale nature, due to the necessity of having materials hand-tagged in the development process.

These factors are intentional design criteria for a system which is not intended to be used for the type of analysis carried out in the present work. It is therefore not a criticism that USAS cannot be used for the analysis intended in this thesis, but is rather an effect of its being aimed at a different set of research questions. However, USAS has recently been used in some ways which are congruent to the analyses undertaken below; these have not been at the required scale of granularity for this work, but have used similar statistical measures and methodologies. These two studies, Rayson 2008 and McIntyre

⁸ See further the Conceptual Integration Theory concept of 'human scale' in §4.iv.5.b below.

and Archer 2010, were published while work on this thesis was ongoing, and have provided useful and successful proofs of the type of methodology and statistical analysis undertaken here, although do not replicate its procedure or outcomes.

Moving away from USAS, the only other semantic taggers currently available operate within the environs of computer-based information retrieval (IR) and are more strictly part of computational linguistics (where linguistics arises from computing needs) than corpus linguistics (which generally take the opposite approach). IR methodology generally involves building a system which identifies broad meanings for later text-mining (for example, finding email addresses, or discourse topics, or temporal expressions (Schilder and Habel 2001)). None involve semantic tagging in the way linguists use the term; many are narrow in their focus (eg to automatically retrieve weather descriptions using a predefined lexicon; Boufaden 2003) and do not attempt to relate semantic tags to semantic domains as a cognitive linguist requires. In fact, they necessarily avoid the granularity required to engage with those domains; for example, Demetriou and Atwell (2001) argue that Roget's *Thesaurus* is 'too fine-grained' with its 1000 categories. None of these approaches are therefore appropriate for the analysis necessary in this thesis.

2.II: METHODOLOGY

2.II.I: INTRODUCTION

It should be by now clear that no currently-available system can be used for the macro-level analysis necessary in this thesis to establish the overall semantic distribution of the texts in a highly-analogical popular science corpus. This section introduces and explicates the technique of semantic profiling developed for this thesis. It firstly describes the data, including the compilation and makeup of the analogical popular science corpus, and then sets out the analytic

process which sets out to address the key research question of this part of the thesis. Concluding with a summary of the procedure, this chapter then gives way to Chapter 3's implementation of the macro-level analysis on the corpus.

This chapter and the next uses two new technical terms which are necessary to describe the directionality of the semantic/lexical analyses: domains are *instantiated* by lexemes, and lexemes *activate* a semantic domain.

2.II.2: THE DATA

Four sets of data are used: the analogical popular science text corpus itself, a second corpus which is made up of comparison texts expected to be non-analogical, a third reference corpus for statistical reasons, and finally, the semantic database of the *Historical Thesaurus of English*, used to provide the link between electronic text analysis and semantic domains. These four are outlined below.

2.ii.2.a: The Corpora

The main corpus assembled for this part of the thesis consists of the full text of four popular science books in the areas of physics and mathematics. This main corpus is used in conjunction with three comparison texts and a reference corpus, for the purpose of analysis.

2.ii.2.a.i: The Main Corpus

The four main texts analysed are Brian Greene's *The Fabric of the Cosmos* (2004), Michio Kaku's *Parallel Worlds* (2005), Marcus du Sautoy's *The Music of the Primes* (2003), and Lisa Randall's *Warped Passages* (2005).

The selection of these core texts was based on five main factors: their content, their authors' credentials, their date of publication, their general reception, and their electronic availability. This meant that, out of the large number of popular science books that have been published, the thesis' focus was narrowed down to the four above using these criteria. In more detail, the content

criterion was that the texts were to be in the fields of mathematics and physics, as described in Chapter I above. Secondly, as a means of ensuring the books' authors were qualified to a high level in the areas of science they discuss, the authors were narrowed to those who are research-active academics in the fields through which they publish. Thirdly, in order to use modern texts which are readily available, the texts should have been very recent at the time this thesis was begun (published in the three years preceding 2006, and thus no earlier than 2003). Fourthly, they should be well-received by the readership in general as successful examples of the genre; this is a notoriously difficult metric to measure, but as a rough proxy Amazon.com reader reviews were chosen (see below), with the criterion being that the books should have a minimum star rating on that site of 4 out of 5. Fifthly, they needed to be available electronically as eBooks.

With regards to the four chosen texts, they all met those criteria unambiguously. *The Fabric of the Cosmos* (hereafter *FC*), *Parallel Worlds* (hereafter *PW*), and *Warped Passages* (hereafter *WP*) centre around various aspects of theoretical physics, while *The Music of the Primes* (hereafter *MP*) is on prime number theory. *FC*'s author, Brian Greene, is Professor of Mathematics and Physics at Columbia University, Michio Kaku (*PW*) is the Henry Semat Professor of Theoretical Physics at the City University of New York, Marcus du Sautoy (*MP*) is Professor of Mathematics at the University of Oxford and Fellow of New College, and Lisa Randall (*WP*) is Professor of Theoretical Physics at Harvard University.⁹ The books all range in publication date from 2003 to 2005. All have an average of 4.5 stars (out of 5) on Amazon.com reviews as of late 2010 (prospective texts with fewer than 10 reviews were discarded as the lack of reviews made it difficult to gain an accurate average-review rating, so that a secondary criterion here was review volume; as of early 2007, *FC* had 240

⁹ Information on all four can be found at their university webpages. These are: Greene, <http://www.columbia.edu/cu/physics/fac-bios/Greene/faculty.html>; Kaku, <http://www.cuny.edu/about/people/faculty.html>; du Sautoy, <http://people.maths.ox.ac.uk/dusautoy/>; Randall, <http://physics.harvard.edu/people/facpages/randall.html> (confirmed 1 November 2011).

reviews, *PW* 70, *MP* 28 and *WP* 161). All were available as eBooks, although in a range of formats; after being purchased they were converted to plain text format for analysis.

In this way, this corpus selection does not aim at comprehensiveness or generic balance, but rather mixes key indicators of relevance to the thesis' research questions (credentials, content, recent publication and positive general reception) with pragmatic factors (availability of content). The corpus is also of a substantial length, large enough for the wide-scale analysis here and not too large to be prohibitive for use in Part II below; the individual texts range in size from 125,000 words to 192,000, with an average of 150,058 words and a total corpus size of 600,233 words.

2.ii.2.a.2: The Comparison Texts

Three comparison texts were chosen, to act as checks on the validity of the process outlined below. In order to be a generally-valid procedure, the semantic profiling technique should work on a very wide range of texts, not just the narrow genre analysed in this thesis. Three other texts were thus analysed, to provide a range of checks against the technique. The first of these is Douglas R. Hofstadter's 1979 *Gödel, Escher, Bach: An Eternal Golden Braid* (hereafter *C-GEB*; the C prefix indicates this is a comparison-only text), a Pulitzer-prize-winning classic in popular science. This meets many of the criteria above (Hofstadter is Professor of Cognitive Science at the University of Indiana, *C-GEB* has a 4.5 star average on Amazon out of 263 reviews, the text is available electronically), but is older than the others and focuses on the nature of logical systems, intelligence, and cognition. It functions as an example of popular cognitive science, not in the fields of physics or pure mathematics, and its selection is based on it being a core part of the popular science 'canon', as evidenced by the prolific reviews and citations listed on the inside cover of any print version of the text. The second comparison text was chosen to be one similar in topic but not in genre, and is W.A. Coppel's 2009 *Number Theory* textbook (hereafter *C-*

NT). It is particularly free of descriptive material, instead being a technical outline, and so fits well as a non-analogical comparison text. The third of these texts was intended to be a contrast in both style and topic, while being one with predictive qualities; that is, a text which the present analyst knows very well and so can predict what its frequently-occurring semantic domains would be, in order to act as a check on any results obtained by the process below. The final comparison text is therefore my 2005 Masters' thesis, *Cognitive-Linguistic Manipulation and Persuasion in Agatha Christie* (hereafter C-AC). It focuses on a stylistic description of selected detective novels of that author, is relatively recent, and is available electronically. More importantly, being my own copyright, it also acts as a text which is distributed online and in Appendix III alongside the programs below, whereas all the other electronic texts are analysed here under fair-use provisions following purchase, and so are not redistributable.

2.ii.2.a.3 The Reference Corpus

Finally, a reference corpus was required with which to compare the semantic distribution of the above texts, and so identify domains used more frequently (and possibly characteristically) in one text rather than in others. As the main texts all focus on expository information for a general audience, it was decided not to use typical English reference corpora such as the Brown Corpus, the British National Corpus (or the BNC Sampler/BNC Baby), as the expository component of these (where identifiable) was too small. A larger reference corpus was needed which was easily available, substantially-sized and with an expository discourse function similar to that of popular science.

The reference corpus therefore consists of a million sentences taken randomly from the English Wikipedia.¹⁰ The Wikipedia project is intended for a general audience, is written in formal third-person English, covers a wide range of

¹⁰ Available online from David MacIver's webpage at <http://goo.gl/gBDI3>. The English Wikipedia is at <http://en.wikipedia.org/wiki>.

topics and is also available under a generous open-use license. These random million sentences thus provide a 14,077,945-word reference corpus of broadly similar expository aims to the core texts above, but without a substantial generic bias (the random nature of the corpus means that even distribution of topics cannot be guaranteed; however the high number of sentences is intended to minimize generic bias, as the reference corpus is around 1.1% of the full English Wikipedia). Importantly for the purposes of the statistical analysis below, the reference corpus is significantly larger than any of the individual texts below (the largest is *C-GEB* at 264,000 words, which is still only 1.9% of the size of the reference corpus).

2.ii.2.b: The *Historical Thesaurus of English*

Finally, the semantic data used in the thesis is provided by the database of *The Historical Thesaurus of English* (published in 2009 as *The Historical Thesaurus of the OED* and hereafter known as the *HT*). The *HT* is a semantic arrangement of the complete contents of the first edition of the *Oxford English Dictionary*, together with the later supplements and the 1990s three-volume *Additions* series (being approximately equivalent to the second edition of the *OED* in 1989 with the later *Additions*), and incorporating *A Thesaurus of Old English* (Roberts and Kay 2000 [1995]). This provides a fine-grained semantic hierarchy of all the recorded senses in English from the Anglo-Saxon period to c.2000CE, within a 12-level system of classification based wholly on the data available in its sources (see Kay *et al* 2009:xiiiiff). In this system, each category of words is nested within other, wider categories, so that, for example, the verb category *Live dissolutely* is within *Licentiousness*, itself adjacent to *Guilt* and *Rascalry* and within the wider category *Morality*. This hierarchical structure differs from the organization of many other thesauri, such as that of Peter Mark Roget (1852, see also Davidson 2004 and Hüllen 2004 and 2009). While Roget's categories exist in a single linear sequence, *HT* categories can relate to others either horizontally (on the same hierarchical level) or vertically (on a higher or lower level, either containing or being contained by another category). In addition, each concept

is able to contain a series of subcategories within itself, separate from the main sequence. It is this complex hierarchical structure which helps make the *HT* database so useful for the purposes of this thesis: while each individual point in the hierarchy contains word entries for the concept it represents, it can also be subsumed within a higher-level category and surrounded by sibling categories of similar meanings for the purposes of abstraction and general semantic-field identification.

The size of the *HT* also makes it amenable to computational analysis. As the largest thesaurus in the world, and the only historical thesaurus for any language, the current version of its database (as of 2011) contains 793,747 word entries, compared to *OED2*'s 616,500 (Algeo 1990:137), all within 236,346 categories, each representing a distinct concept. Taking into account each field stored within it, the database itself contains approximately 22.7 million pieces of data.¹¹

These figures represent the full database. However, for the synchronic purposes of this thesis, the diachronic dimension of *HT* is significant not for its historical information, but in its possibilities for filtering the full thesaurus in order to create a solely modern subset of the data. Delineating the idea of 'modern' does provide some complex issues to be clarified, though, and these issues are addressed below. It is clear that including non-current words in the data will add to any spurious results that the procedure risks, for example by including obsolete homonyms which will be misidentified as instantiating domains not intended by the author or acknowledged by the reader. The nature of the *OED* data on which *HT* is based, however, makes identifying non-obsolete (and therefore current) word forms somewhat tricky. *OEDI*, known at the time as *A*

¹¹ There has not been much opportunity for research has thus far into the *HT*'s potential for computational analysis, given the recent nature of its completion. For those studies that have been done, mainly at conferences, see Alexander 2010a, 2010b, 2011a, 2011b, Alexander and Kay 2011a, 2011b, Alexander and Struan in press, Alexander and Wild 2010, Anderson *et al* 2011a, 2011b, Archer and Kay 2010, Archer 2011, Johnson *et al* 2010, Juuso *et al* 2011, and Kay and Alexander 2010.

New English Dictionary on Historical Principles (Murray *et al* 1884-1928), was published between 1884 and 1928. A 1933 *Supplement* (Craigie and Onions 1933) added antedatings and omitted words, mostly from the first half of the alphabet, and then a four-volume *Supplement* series (Burchfield 1972-1986), mostly covering new words coined in the period following *OED1* data collection, was produced in the 1970s and 1980s. All these merged together, with a small number of minor changes, formed the ‘revised’ *OED2* (Simpson and Weiner *et al* 1989), which was later itself supplemented by the three-volume 1990s *Additions* series (Simpson and Weiner 1993, two volumes, and Proffitt and Simpson 1997). All this means that swathes of the *OED*, particularly in the first half of the alphabet, remain unrevised and – more importantly for the purposes of this thesis – uncited from the late 1800s. The in-progress *OED3* revision (Simpson and Weiner *et al* 2000-) will correct this, but for now the nature of the *OED* data makes it difficult to automatically tell if, say, a word beginning with the letter A and with its last citation evidence in 1870 is in fact a Victorian archaism or a current word form which has not been supplemented or revised since *OED*’s first publication in 1888. *HT* policy in these matters has been to consider a word current if it appears in the *Concise Oxford Dictionary* (Fowler and Fowler 1911, Sykes 1982 and Soanes and Stevenson 2009), and there is a field in the database which marks, in a yes/no fashion, if it can be considered current. This current marker is thus based on one of three factors: either *OED* citation evidence, particularly for those words appearing in the *Supplements* or *Additions* and last cited in the previous few decades; or the evidence of the *Concise*; or the knowledge of an editor. Post-1888 terms are therefore highly problematic with regards to their identification as current. For the sake of as wide an inclusion as possible, particularly for technical items where currency is even more difficult to determine due to low frequency of appearance in citations and the unlikelihood of its appearance in a concise desk dictionary, this thesis follows other work (eg Alexander 2011a) in defining ‘modern’ as either marked as current in the *HT* or with a final citation date in

the past 150 years. This may include a small number of obsolete words, but avoids the risk of reducing the dataset in technical areas beyond what can reasonably be termed non-current.

Therefore, when appropriately filtered, the *HT* data yields the largest thesaurus of contemporary English currently available, arranged within the complex hierarchical structure outlined above. Viewing this modern data as a series of word forms instantiating conceptual domains, with these domains themselves being defined along a cline of granularity and semantic precision in the *HT* classification hierarchy, gives a dataset which is ideally suited for the analysis attempted in this thesis.

2.ii.2.c: Acronymic Summary

This chapter has introduced a large number of acronyms, which are summarized below for the benefit of readers.

FC: The Fabric of the Cosmos: Space, Time and the Texture of Reality, Brian Greene.

PW: Parallel Worlds: The Science of Alternative Universes and our Future in the Cosmos, Michio Kaku.

MP: The Music of the Primes: Why An Unsolved Problem in Mathematics Matters, Marcus du Sautoy.

WP: Warped Passages: Unravelling the Mysteries of the Universe's Hidden Dimensions, Lisa Randall.

C-GEB: Gödel, Escher, Bach: An Eternal Golden Braid (Comparison Text), Douglas R. Hofstadter.

C-NT: Number Theory: An Introduction, Second Edition (Comparison Text), W.A. Coppel.

C-AC: Cognitive-Linguistic Manipulation and Persuasion in Agatha Christie (Comparison Text), Marc Alexander.

HT: The Historical Thesaurus of English (2009 print version: *The Historical Thesaurus of the Oxford English Dictionary*), Christian Kay, Jane Roberts, Michael Samuels and Irené Wotherspoon (eds).

2.II.3: THE PROCESS

As outlined at the start of this chapter, the aim of the computational technique described in this part of the thesis is to take a text and give a resulting list of all the semantic domains activated in this text. A few assumptions were made in the process, which are explained in later sections. In outline form, a set of Python programs were written using a series of routines known as the Natural Language Toolkit (NLTK; Bird *et al* 2009), which take the text fed into it and reduce the words in that text to their lower-case base lemmas (sometimes known as their dictionary form; *run*, *Runs*, *running* are all reduced to *run*, while *Cat* and *cats* are reduced to *cat*). Then frequently-occurring closed-class words are excluded on the basis of a pre-existing list of these words (thus eliminating, for example, *the*, *a*, *their*, *some*, and so on). The number of times each word occurs in that text is counted, and then the table of lemmas and the count of each is run through the *HT* database. Each time a word in that table matches a word in the *HT* database, it counts as an activation of that domain (for a discussion of the problems of polysemy that arise from this, see §2.ii.5 below) and the number of times each domain is activated is recorded in a new table, containing all the *HT* categories and the counts of each.

In order to find those domains which are activated unusually frequently in comparison with the reference corpus, a statistical measure is used based on the keyword. The SQL formulae for calculating these were based on Paul Rayson's log-likelihood ratio formulae for Microsoft Excel.¹² Given that the activation of

¹² From <http://ucrel.lancs.ac.uk/llwizard.html>. The original source of the log-likelihood formula is Dunning 1993.

semantic domains is non-random in a text, the log-likelihood ratio (LL hereafter; sometimes referred to solely – and in a more technically accurate manner – as *likelihood ratio*, eg in Meyer 2002:165) formula is a good choice for establishing significance here. The LL process identifies this statistical significance, but an analyst is still required to decide if the key item is significant with respect to the nature of the text or a predefined research question. For example, if the C-AC comparison text has 172 references to HT category 01.02.02.04.01.01 *Murder*,¹³ and the reference corpus has 8714 references to it, then the first step is to provisionally assume that both texts should have identical proportions of words activating that category (this is known as the null hypothesis; it is the default position one might expect should the category *not* be key to the text, and so exists to be disproved by the statistical test). The expected frequency of that category in both the text and the reference corpus is calculated given the expectation that its frequency *should not be different* (that is, they should both be made up of the same proportion of categories), which in this case gives a prediction in the text of 9.8 occurrences and in the reference corpus of 8876.2 occurrences. Comparing these to the actual occurrences (172 and 8714 respectively) we find that the reference corpus matches the prediction fairly closely, but that the text does not, and in fact is out by an order of magnitude. The LL formula takes these two deviations (the difference between the prediction and the actuality) and returns a score of significance; as McIntyre and Walker put it, ‘the higher the LL value the more key or statistically significant an item is or the higher the likelihood that the unusually high or low occurrence of an item is not due to chance’ (2010:517). In the case of this example, the LL value is 664.2, which is an extraordinarily high figure. Rayson *et al* 2004 suggest figures above 15.13 to be significant, while figures above 37.3 mean there is a probability of only one in one billion that the difference is not

¹³ The long number is the HT hierarchical reference code. It locates this category within 01.02.02.04 *Killing*, itself within 01.02.02 *Death*, within 01.02 *Life*, within 01 *The Physical World*. Each category inherits the reference number of its parent category with an additional final number group, this final number increasing with each new category in that section.

statistically significant. The LL value of 664, therefore, suggests quite strongly that the semantic deviation here is significant, and this is borne out by the data: the *C-AC* text, being a thesis on the stylistics of detective fiction, mentions murder a great deal. The highest LL figures for this text include in the areas of *Rhetoric* (*HT* 02.08.03.06.02; LL 2074.1), *Suspicion* (*HT* 02.01.13.08.02; LL 1477.4), *Reader* [=one who reads] (*HT* 03.08.09.01; LL 1183.9), which are all to be expected by anyone who knows that text. Although the LL test is an established one (for a range of users see, *inter alia*, Baayen 2008, Baker 2010, Larose 2006, and Mendoza-Denton *et al* 2003), the fact these significant domains match with predictions of the domains I knew I used while writing that text acts as a useful and important check on the validity of the procedure.

So while this procedure identifies semantic domains, it gives only lists of domains, statistical relevance and frequency distributions. While these are useful figures, for stylistic analysis it is important to be able to go back to the text and see which words purportedly activated that domain and where they occur in the text. Firstly, there will be significant amounts of ‘noise’; this is a standard term for unwanted computational identifications based on non-optimal routines, and consists in this thesis of mis-identified domains due to the somewhat intractable dual problems of polysemy and homonymy in English. To identify and discount this noise, it is necessary to look at the resulting key domains and remove those which are there due to the interference of homonymous or polysemous terms. This can easily be done by an analyst examining the lexical spread of the domain and how many items there are. For lexical spread, it is important to examine how many different lexical items activate the particular domains in question; the fewer there are, particularly with reference to the size of the *HT* category, the more likely it is that lexical interference is taking place. Similarly, if the lexical items themselves are polysemes/homonyms which are also available in a domain which is relevant to the literal topic of the text in question, then with the exceptions of puns or other wordplay, there is a greater chance of the domain being ‘noise’.

Data giving the lexical-domain relationships of the texts in question is therefore required. This was accomplished by means of another Python program, similar to the one above, which reduces texts to their lemmas and removes their function words and punctuation, before returning a numbered list of each word in the resulting filtered text (so that, for example, *The cat sits on the Mats [sic]* would be returned as *1 cat, 2 sit, 3 mat*, with each word and number pair on a new line). These versions of the text (called the Ordered List versions) were then put through the same database query as above, giving each semantic domain potentially activated by each lemma (eg *1 cat 01.02.03.04, 1 cat 02.03.04.05, 1 cat 02.06.07.08, 2 sit 01.02.01.02* etc). When viewed as a list, and then sorted or filtered based on the domain number, this data shows the activating lemmas of each domain alongside their positions in the text.

With these two datasets (the LL-scored domain list and the lemma-based Ordered List), the next stage of analysis constitutes identifying the key domains for each text and then filtering them to eliminate noise and non-analogical uses. For this, the hundred semantic domains for each text with the highest LL score (the ‘key domains’) are taken and tagged according to whether they are or are not consistent with or relevant to the *literal* topic of the text. This judgement is a decision which needs to be made by the analyst, and therefore has some potential for problematic decisions, so following some experimentation the following guidelines were established:

1. The single domain considered most representative of the text (eg *01.04.07 Physics*) is automatically considered relevant. (This will be the domain with the highest LL, by definition. Should it not be, the analyst is required to explain this.)
2. The neighbours of this domain (that is, within the same level-2 category; so for *01.04.07 Physics* this would be the remainder of *01.04 Matter* and its daughter categories) are considered relevant unless this domain is so large or so varied as to make this judgement inadvisable, at which point

the neighbours are only considered likely to be relevant with regards to the analyst's judgement.

3. Nearby domains to the core domain (eg *01.05.07 Space* near to *01.04.07 Physics*) are also considered to be relevant should they be, in a strictly literal interpretation, semantically appropriate in the analyst's judgement. This includes their daughter categories. This is, however, not applied should the domain be an abstract one (eg *01.05.07.05.25.05 State of being scattered/dispersed*), as abstractions are often used analogically.
4. Areas of immediate semantic importance to the contextual situation within which the text sites itself are considered relevant. For example, the texts are all academic-related, so the high LL of *03.06.06.02 College/university* is to be expected (score from 500 to 4,906, with the exception of *C-AC*), as is *01.01.06.10.01 United States*, where the majority were written and from which a lot of research originates (LL score from 711 to 1,836, excluding *C-AC* and *MP*, which are British texts). There are some items which are also standard participants in any discourse (eg *01.02.07.08.04 Person*, LL from 401 to 2,326). This criterion was applied narrowly, with a minimum of speculation and based only on available paratextual information such as the cover blurbs, the author biographies, and, in harmony with the selection criteria in §2.ii.2.1.a above, and the book summaries on Amazon.com (eg *03.04.06.13.08 A/the government*, LL from 512 to 1449 with the exception of *C-AC*, could have been identified as the likely source of funding of research reported on, or possibly the owners of laboratories etc, but was categorized as non-relevant as the link was not immediate and requires information from the text rather than the paratext).

The remaining domains were considered non-relevant with regards to the literal topic of the text, and were thus taken forward for analysis. Of these non-relevant domains, each was then checked for 'noise' (accidental identification,

usually on the basis on homonymy). This requires more analyst judgement than above, and probability of a key domain being noise is based on the following core criterion:

- A literal domain with a low lexical spread and where that lexical spread contains polysemous words also found in the text's key relevant categories has a high probability of being noise. This probability increases as the lexical spread thins and also as the LL score of the alternate-relevant category rises.

This checking was done over the Ordered List versions of the texts, filtering the full list by the *HT* category in that field. This procedure finds a list of all the words which activate that domain, alongside where in the Ordered List it comes. The lexical spread can be identified easily by browsing the list, but for long categories the query can sort, group and summarize the data by word.

The actual domain analyses are rather more qualitative than quantitative, and are outlined in Chapter 3.

2.II.4: SOME TECHNICAL DETAIL

This brief section contains some technical summaries that readers uninterested in the mechanics of the programs described above may not wish to read in detail.

Following some limited editing of the texts (removing prefatory matter, tables of contents, bibliographies, indices, page numbers and paratext, but not removing glossaries where one exists (eg in *PW*), as glossaries have expositional content), the technique begins with custom Python programs designed to take a text, tokenize its contents, filter the token list based on a pre-defined stopword corpus, and then create a frequency distribution of that wordlist.

These Python programs use the NLTK (see Bird *et al* 2009) and were originally used on Python 2.6.4, but have been tested subsequently on Python 2.7. Then SQL queries create a datasheet linking those lemma frequency distributions to

HT categories, treating all lemmas as activating each possible referent simultaneously (see §2.ii.5 below for an explanation of why this was considered acceptable). The resulting semantic frequency distributions for the text at hand and the reference corpus are compared using a log-likelihood statistical test similar to those used in Wmatrix (Rayson 2009) and WordSmith (Scott 2008), giving a measure indicating the statistical significance of the deviation in frequency, based on Paul Rayson's Microsoft Excel formulae, and incorporating an embedded if-formula to ensure that the text section of the domain activation table is not null. The datasheet created in this manner is then exported to csv format and analysed in Microsoft Excel. Each text in this dataset then has its hundred highest-scoring LL domains examined as outlined in §3.i below.

The second dataset, one listing lemmas in order of use in the text alongside each lemma's activated domains, is accomplished by a further Python program which is identical to the one above but, after the filtering process, instead returns a list of lemmas alongside a simple count of where the lemma is in the text. This list is put into a similar SQL query, returning multiple rows of the lemma list with every domain activated by each lemma. This dataset is used to qualitatively filter noise from the domain listing above. Additional SQL statements are added to this select query as needed to filter the datasheet to show only those domains under investigation.

2.II.5: POLYSEMY

The technique used in this thesis takes, relatively speaking, a brute force approach to semantic profiling with regards to the somewhat intractable problems of polysemy and homonymy. It, for the purposes of expediency, assumes all words are polysemes rather than homonyms, and that lexical items activate in the reader all the possible meanings of the item under question. This is not psychologically or cognitively plausible, but is necessary for the purposes of the current analysis. The procedure is acceptable in this way because the

process used in this thesis is not technically that of a tagger, but rather of a wide-scale identifier. Therefore, its quantitative analysis introduces some analytical complexity which an analyst is required to deal with qualitatively. This complexity could be viewed as a technical failure, with the aim of domain identification (such as semantic tagging) being to narrow the possible meanings of lexical items under consideration to a single meaning, which is intended to represent the meaning of that lexical item as used at its contextual situation in the text. There are three theory-driven objections to this approach which arise when undertaking the work in this thesis:

1. It is not systemically simple to identify lexical items being used with multiple simultaneous meanings, such as those used as poetic metaphor, puns, or, with particular relevance to this thesis, explanatory analogy. The possible monosemy/polysemy dichotomy is therefore more accurately described as a cline. Artificially imposing a single meaning on a cline is a sub-optimal solution to a complex problem.
2. In most of cognitive linguistics, word meaning is understood as network-based (see *inter alia*, Evans 2009, Fauconnier and Turner 2002 and Hudson 2007). As such, meaning activation is a plural and protean process, giving a 'spreading' effect wherein activation of a given node 'spills over' (Reisberg 1997:256ff) to nearby conceptual nodes. This, part of the core assumptions of the model of cognitive linguistics followed in the second half of this thesis, is not easily reconciled with the one-lexeme-one-meaning view, and this second view is therefore not followed herein.
3. Finally, in the dominant paradigms of modern psycholinguistics (eg Pederson in Geeraerts and Cuyckens 2007a), meaning is strictly understood at its core as being idiolectal, and so ambiguity, even if not intentional, is not easily resolvable independent of an interpreting mind. Modelling this within the context of a shared language-system for

computational purposes is a complex (although inevitable) problem which is far outwith the scope of this present thesis.

Therefore, the process used here aims to be maximally-inclusive in the data it provides to the analyst, incorporating excess noise as the price to pay for not risking overzealous or inaccurate methods of narrowing focus. Ignoring the obvious choice of narrowing the data based on grammatical categories is one such example: were there to be an available part of speech tagger of substantial reliability, this would be a plausible method of reducing data noise. But given the subject matter of the texts at hand and the relative lack of accuracy of current taggers, this possibility was abandoned. Maximal inclusivity has its flaws, but providing excessive data and then working on more sophisticated methods for an analyst to access and comprehend that data was felt to be a preferable aim over narrowing the data and accepting the loss of breadth that avenue would entail.

2.II.6: RATIONALE AND ALTERNATIVES

In this section, two methodological questions are addressed with regards to the existing availability of alternatives: why the Python language, and why the *HT*.

The Python programming language is used widely for natural language processing and has an established set of routines for language work, the Natural Language Toolkit (NLTK).¹⁴ It is somewhat rarely used within corpus linguistics, but is increasingly popular in computational linguistics. Its lack of use in more traditional corpus linguistics is perhaps because it has no graphical user interface and must instead be used at the command line or in a text-based integrated development environment (IDE) such as Python IDLE.¹⁵ This is

¹⁴ The main Python site is at <http://www.python.org>, where the language and its IDE can be downloaded. Note that the work in this thesis uses Python 2.6.3 and has been tested on subsequent 2.x releases; the programs here will not run on Python 3.x, which is backwards-incompatible with 2.x.

¹⁵ IDLE has no official acronymic derivation, but is rather a pun on IDE referencing Eric Idle from the comedy group Monty Python.

unfortunate, as the language is ideally-suited to corpus analysis, and the NLTK even comes with preset copies of 74 different corpora¹⁶ and a freely-available online book explaining its use for a general, non-programming audience.¹⁷ Python's philosophy is to emphasize the readability of code, and a non-programmer could easily pick up the whole scope of the NLTK within a week or so, making the methodology here easier to replicate. Python and the NLTK are both free and open-source, and so available for anyone with an internet connection to download and use.

In terms of language choice, the Perl language would be a legitimate alternative; the use of Python over Perl (or, for that matter, Ruby or Java) is one of programmer preference and ease/speed of writing. Similarly, other users may find an R implementation of the statistics analysis more accessible than the lengthy SQL query structure used here, which would be entirely acceptable (cf, *inter alia*, Baayen 2008 or Gries 2009). Neither change would substantially alter the process here other than in details of implementation.

More importantly, though, Python and the NLTK are attractive as they are powerful tools for experimental computational linguistics and rapid development in the digital humanities: the programs written for this thesis involve complex operations (such as tokenizing, filtering, lemmatizing and then creating a frequency distribution of the 14-million word reference corpus) and the Python routine developed could complete this and then write the results to a CSV file in a reasonable timespan. For example, a relatively-underpowered laptop, designed for portability rather than computing power, did this in under eight minutes; a more recent desktop machine completed it in four. Other routines on smaller texts, such as the books used in this thesis, are often completed within 30 seconds or so. This allows a non-expert programmer

¹⁶ These corpora include the Brown Corpus (in both text and TEI-XML), selections from Project Gutenberg, US Presidential Inaugural Addresses, a wide set of English names, stopwords, Reuters news articles, WordNet, a multilingual parallel corpus from the European Parliament, and others. See http://nltk.googlecode.com/svn/trunk/nltk_data/index.xml for more details.

¹⁷ See <http://www.nltk.org/book>, also printed as Bird *et al* 2009.

to test a program, evaluate its results, alter the code and re-test the results iteratively until the program succeeds in its purpose. Thus, NLTK is an important tool for experimental corpus linguistics as it allows rapid development and testing of untried routines.

In terms of pre-written packages available, there are many alternatives to Python which would be acceptable for many users. For example, WordSmith Tools for Windows (Scott 2008) is a widely-used software package, and is ideal for certain corpus linguistics work with established routes (eg keyword analysis of a corpus, or creating general word lists), but is a commercial piece of software and has only limited customisable functionality for non-standard corpus work. These limitations made it unsuitable for use in this thesis. As a matter of policy, I also wished to use free and open-source methodologies wherever possible rather than commercial paid-for and platform-dependent solutions. The freeware AntConc software¹⁸ is similarly restricted in its usability, although not in its cost or platform.

There is a final pedagogical point here, in that a case could be made that it is preferable in a thesis for the author to create programs him or herself if possible. A solid engagement and knowledge of the digital analysis process at hand, including how it operates, each step it takes, its assumptions, statistical measures, pitfalls and any possible errors introduced is a useful requirement for an analyst who wishes to create new procedures and evaluate their usefulness. This is perhaps not a strong point to be made in the case of established and well-known routines and analyses such as collocational studies, keyword analysis, and so on, but for the present thesis programming the technique from scratch where possible was considered an important stage.

Therefore, Python and the NLTK were chosen as the development environment for the thesis, with the *HT* stages accomplished by use of SQL queries running on the *HT* database. These queries were executed in Microsoft

¹⁸ <http://www.antlab.sci.waseda.ac.jp/software.html>

Access, this being the Historical Thesaurus project's preferred storage format. It would have been possible to easily convert this data into XML and thence into a Python dictionary hash, but as the *HT* database is under constant development and alteration, it was considered preferable to work on the Access version so as to use the most recent data possible. The use of SQL as the query language as opposed to the Access visual query builder is an attempt to make the process as clear and reproducible as possible, at the expense of ease of use.

The Python programs and SQL queries are included as Appendix I. Alongside these are annotated copies which explain the purpose of each line. All are also available for download at <http://phd.seldomend.com> under an open-source Apache 2.0 license.

Turning to the *HT*, its superiority over other systems (for example, the USAS tagger or WordNet) has been outlined above. An important further reason is that of my own access to the dataset; as research and editorial staff on the Thesaurus since 2005, I have a great deal of experience and had pre-publication access to its full dataset, in addition to experience in editing material and designing the latest version of its database. As a result, I have detailed knowledge of the structure and use of the dataset, and in its potential to address the research questions raised in this thesis.

2.III: METHODOLOGICAL OVERVIEW

The process outlined above describes an iteratively-tested semantic domain identification system across large spans of text. It combines established statistical techniques with custom-written programs and new datasets to allow this part of the thesis to test the hypotheses outlined in §1.i above with regards to the systematic nature of the use of analogy in popular science.

To conclude, it is important to emphasize that the analysis presented here is a limited computational attempt at identifying semantic domains in use across

long and complex texts. It is not, and was not intended as, a semantic-tagging system. This caveat has been discussed above as a limitation of analysis due to the scope of the present thesis, although it is certainly computationally possible, particularly with modern computational resources. A semantic tagger would, however, necessitate the use of probabilistic contextual models, narrowing possible semantic alternates with reference to the extended, progressive high-certainty discourse neighbourhood of each lexical item in question, and would require also to contain an iterative part-of-speech tagging model, again probabilistically-based and using the full range of formal, functional and semantic information (ideally alongside large banks of n -gram data) to assign syntactic and corresponding semantic labels. It is possible in principle for this to be done using Python, the programming language employed in the wide-domain-analysis process described throughout this chapter. The development of a systematic, data-driven, hierarchical, diachronic syntactic-semantic tagger is a necessary priority for English corpus linguistics (see §6.iii), and this thesis aims to have made a small contribution towards demonstrating the *HT*'s potential in this area.

Chapter 3 below applies the techniques outlined in this chapter to both the core corpus and the comparison corpus.

CHAPTER 3: KEY ANALOGICAL DOMAINS

3.1: INTRODUCTION

The analysis undertaken here applies the methodology of Chapter 2 to address the following research questions:

1. Does the technique described in Chapter 2 identify key domains in the corpus of popular science texts?
2. Does the technique identify key analogical domains in such texts?
3. What can be identified about the stylistic use of such key analogical domains across the whole text, and what, if anything, characterizes this use?

This chapter therefore forms the core of the macro-level first part of this thesis, by testing these hypotheses and establishing the extent of analogical content in the corpus. Once this has been established, the second part of this thesis, following this chapter, addresses questions of local textual usage of such analogies, and of the semantic and stylistic construction of analogical meaning at the small-scale discourse level.

In terms of structure, the chapter proceeds through an analysis of each text in turn (*FC*, *PW*, *MP* and *WP*) and then compares this analysis with that of the comparison texts (*C-GEB*, *C-NT* and *C-AC*) as a means of contrasting the results and so demonstrating that the results presented here with regard to the core texts are, at least in part, due to their analogical nature. The analysis of the first text, *FC*, describes the process of identifying key domains in more detail than the others, as a matter of methodological illustration. The following analyses use the same method as in *FC* and Chapter 2 above, but shorten its exposition somewhat. The length of §3.ii.1 below is therefore substantially longer than comparable sections in other analyses.

Each analysis proceeds as follows: The text is taken and each key domain is analysed in turn, grouped by the three major divisions of the *HT*. Each domain is analysed for relevance against the criteria in §2.ii.3 above, and then the key literal and analogical domains are separated and listed. These sections and decisions are explained in detail below.

3.II: ANALYSIS: CORE TEXTS

3.II.I: *THE FABRIC OF THE COSMOS*

The hundred domains within *The Fabric of the Cosmos* with the highest LL ranged from a score of 13,655 to 1,307, all an order of magnitude larger than that required (for the key *p*-value of 0.05). 49 of these were considered relevant to the topic of the text, following from the core domain of 01.05.07 Space (the highest LL at 13,655). Of the remaining 51, one third came from the physical world domains (*HT* section 1), 16% from the mental world (section 2), and 51% from the social world (section 3). This section investigates those key domains (here listed in order of LL score) and proceeds to identify those used as analogies throughout the full length of the text.

3.ii.I.a: *FC* Key Domains

3.ii.I.a.i: The Physical World in *FC* Key Domains

Seventeen of the key domains of *FC* are in section 1 of the *HT*: *Action/process of spraying; State of being scattered/dispersed; Sinew/tendon/ligament; Fact of being unoccupied; theory of similarity; Smallness of quantity/amount/degree; Arrangement in row(s)/line(s); Physical sensibility; woven; Seeing/looking; Bodily movement; Mutual relation of parts to whole; Hanging/suspension; South; Nation/nations;*

Revolution/rotation; and *Viscosity*.¹⁹ Of these, the majority are either noise or relevant domains, with a few exceptions.

- *01.04.05.08.03 Action/process of spraying* is noise due to the high occurrence of the word *space* in the text. (See above, §2.ii.3, for a description of ‘noise’ as it is used in this thesis.) It has no lexical spread (that is, no occurrences of multiple lexemes instantiating a particular domain) beyond that single lemma.
- *01.05.07.05.25.05 State of being scattered/dispersed* is also noise due to the word *space*. It, however, does have further lexical spread, but not at a level to re-establish keyness.
- *01.02.05.13.04 Sinew/tendon/ligament* is noise due entirely to the word *string*.
- *01.05.07.04.04.01 Fact of being unoccupied* is noise due to the words *space* and *blank*.
- *01.04.02.23 theory of similarity* is a relevant domain, covering atomic chemistry.
- *01.06.06.07 Smallness of quantity/amount/degree* is a combination of relevant domain and noise due to the words *quantum* and *light*. There is an argument to be made that this is a relevant domain with dead metaphors rather than noise, but this is not pertinent for the purpose of this thesis.
- *01.05.07.05.25.02 Arrangement in row(s)/line(s)* is noise due to the words *rank* and *string*.
- *01.03.00 Physical sensibility* is characterized by terms such as *feel*, *sense*, and *feeling*. It has a wide lexical spread and is not, *prima facie*, literally

¹⁹ Note that key domains which form a main-series category of the *HT* have an initial capital letter, whereas subcategories do not.

relevant to the text. It may be due to the text's representations of scientists working or the well-established almost-dead metaphor of UNDERSTANDING IS TOUCHING; it is thus tentatively here identified as a key analogical domain for analysis in a later section.

- *01.02.09.14.01.03 woven* is made up of a spread of terms such as *fabric, string, stitch, woven, mesh, and loop*. It is neither noise nor literally relevant to the text, and so is a key analogical domain.
- *01.03.07.03 Seeing/looking* is a relevant domain with such terms as *experience, see, and observation* to do with scientific method.
- *01.05.08.01 Bodily movement* is noise due to the words *motion* and *light*.
- *01.06.07.05 Mutual relation of parts to whole* is made up of terms such as *fabric, texture, framework, and tissue*. With some overlap with *01.02.09.14.01.03 woven*, it is neither noise nor literally relevant to the text, and is a key analogical domain.
- *01.05.07.05.09.01 Hanging/suspension* is made up of terms such as *thread, string, flowing, and pendulum*. It overlaps with *01.02.09.14.01.03 woven* and *01.06.07.05 Mutual relation of parts to whole*, and similar to them is a key analogical domain.
- *01.01.03.01.02 South* is a relevant domain of location.
- *01.02.07.08 Nation/nations* is noise due to the word *state*.
- *01.05.08.02.05 Revolution/rotation* is characterized by terms such as *turn, revolution, and spinning* and is therefore a relevant domain.
- *01.04.04.01.03 Viscosity* is made up almost entirely of *string* and *rope*, and is added to the other overlapping key analogical domains above.

There are therefore four overlapping key analogical domains to do with string and fabric, and one further physical sensibility possible analogy for investigation.

3.ii.I.a.2: The Mental World in *FC* Key Domains

Eight of the key domains of *FC* are in section 2 of the *HT: Understanding; Answer, reply; Indo-Hittite; A part of speech; Speculation; Wishing; Faculty of imagination; and Evaluation, estimation, appraisal*. Of these, a few are clearly noise and the remainder are misidentified relevant domains:

- *02.01.08 Understanding* is not noise, but is a relevant domain. It has a very wide lexical spread and few polysemes which overlap with other key domains. It is used throughout the text to refer to the limits and qualities of scientific knowledge.
- *02.01.15.05 Answer, reply* is a further relevant domain, as it contains lexical items such as *answer, yield, explanation, key*, and so on.
- *02.08.02.06 Indo-Hittite*, one of the strangest of the ‘noise’ categories, is an understandable one on investigation: the *Indo-Hittite* category, under *02.08 Language*, contains all the noun and adjective forms of the languages which arose from the Indo-European family, and these overlap with the adjectival forms of country names (*French, Latin, English, Welsh, German, Greek, Yankee* [English], *Jamaican* [English], and so forth). As a consequence, this category appears frequently due to national attributions in the texts. It can be safely ignored in the texts under analysis in this thesis.
- *02.08.04.01.01 A part of speech* is also noise due to the high occurrence of the word *particle* in the text.
- *02.01.13.05.01 Speculation* is a further relevant domain, as it contains lexical items such as *theory, hypothesis, theoretical*, and so on.

- 02.05.03.03.01 *Wishing* is noise, due to the modal verb *would*. It has further lexical spread, but not at a level to re-establish keyness.
- 02.01.07.05 *Faculty of imagination* is a further relevant domain, as it contains lexical items such as *imagination*, *idea*, *think*, and *conception*. As above, it has further lexical spread but not at a level to re-establish keyness.
- 02.01.15.07.03 *Evaluation, estimation, appraisal* has a wide lexical spread, but is a relevant domain.

There are therefore no analogical key domains for *FC* in the Mental World section of the *HT*. This is not a disappointment; we would expect analogies which explain physics in terms of the human mind to be rare. A negative result here is therefore useful in terms of verifying the target domains of such mappings.

3.ii.1.a.3: The Social World in *FC* Key Domains

Twenty-six of the key domains of *FC* are in section 3 of the *HT*: *text of advertisement; end of shift; space for writing; A/the state; type-setting machines; Stringed instruments; Puppetry; a brotherhood; Town; Pattern/design; Bread; A land/country; City; Proportion of notes/rhythm; construction; wage-system; Notation; Politics; metaphor; Tuning/intonation; a room for study; A/the government; Society/the community; An administrative division of territory; Arrangement/appearance of printed matter; and Claim at law*. Of these many are noise but a number are *prima facie* key analogical domains.

- 03.08.05.07.02.02 *text of advertisement* is noise due to the proliferation of the word *space*.
- 03.08.11.02 *end of shift* is also noise due to *space*.
- 03.08.07.06.03 *space for writing* is also noise due to *space*.

- 03.04.06.16 *A/the state* has a limited lexical spread and is made up almost entirely of *region* and *state*. *State* can be disregarded as relevant, with regard to the core condition sense of this lexeme, but *region* is not. Ignoring *state* means that it is difficult to tell if this domain is key or not; however, if the concept lexicalized by *region* is indeed key, it should also appear in other, related domains, which is indeed the case (see below).
- 03.08.08.15 *type-setting machines* is noise due to *space*.
- 03.II.03.01.08.02 *Stringed instruments* is made up almost entirely by the lexeme *string*. Its limited lexical spread makes its identification as key problematic, but the use of *string* in the Physical World above is enough to make it tentatively established as a key analogical domain, its applicability to be decided by later qualitative analysis.
- 03.II.03.04.03.09 *Puppetry* is similar to 03.II.03.01.08.02 *Stringed instruments* above; made up almost wholly of the lexeme *string* it is nonetheless tentatively noted as key here.
- 03.01.04.04.01 *a brotherhood* has a very wide lexical spread, made up of such words as *collective*, *network*, *institute*, *league* and others. It is a relevant domain, with scientists working at Institutes and in networks.
- 03.02.07.01.04.03 *Town* is noise, made up of the words *city* and *satellite*. The *satellite* lexeme is noise because this term for scientific equipment is also used for dependent towns surrounding a city, whereas *city* is a relevant domain to do with location (and also in the titles of universities such as City University of New York, and so forth).
- 03.II.03.02.09.14 *Pattern/design* has a fairly broad lexical spread, made up predominantly by *pattern*, *figure*, and *wave*. *Wave* may well be a technical term, but in combination with the other explicitly non-technical terms, this domain is most likely a key analogical domain.

- 03.07.04.15.07.01 *Bread* is noise due to the lexeme *particle* (a religious term for a portion of antidoron/host; this particular bread section is within 03.07 *Faith*, not food and drink).
- 03.02.07.01.01 *A land/country* has lexemes such as *great, settled, foreign, home, exotic* and *new* (in the sense of being recently settled). While *new* is possible noise, the lexical spread here is enough to provisionally label it as key analogically.
- 03.02.07.01.04.02 *City* is a relevant domain (see also 03.02.07.01.04.03 *Town* above).
- 03.II.03.01.01.04.01 *Proportion of notes/rhythm* is noise due to the polysemous lexemes *time* and *measure*.
- 03.07.04.03.08 *construction* is made up almost entirely of the term *fabric*. Despite its limited lexical spread, it is here key analogically given the use of this lexeme above.
- 03.I0.13.19.08.03 *wage-system* is noise, consisting of the literally relevant *time, relativity* (an economic term meaning relative difference in salaries), and *scale*.
- 03.II.03.01.05.01 *Notation* is noise made up mostly of the word *space*.
- 03.04.06.17 *Politics* is noise made up of the lexemes *state, internal, external* and *alignment*.
- 03.II.03.03.05.II.01 *metaphor* is, sadly, noise made up of the poetics term *figure*.
- 03.II.03.01.01.03.01 *Tuning/intonation* is made up mostly of the word *string*, and, as above, is tentatively marked as key analogically.
- 03.06.06.03 *a room for study* is a relevant domain, made up of terms like *school, study, hall* and *college*.

- 03.04.06.13.08 *A/the government* is noise due primarily to the metonymies *Washington* (used in this domain in senses like ‘Washington has decided to ban pesticides’ but used in the text to refer to places), *Moscow*, and so on.
- 03.01 *Society/the community* is noise due to polysemes such as *general*, *world*, *set*, *circle*, and others.
- 03.04.06.14.04 *An administrative division of territory* is mostly made up of the term *region*, and so is a key analogical domain just as 03.04.06.16 *A/the state* above is. There is some slight noise from the term *hundred* (as in the Chiltern Hundreds).
- 03.08.08.10.01 *Arrangement/appearance of printed matter* is noise due to *space* once more.
- 03.04.13.12.06.01.01 *Claim at law* is noise due to the technical term *quantum*.

There are therefore eight key analogical domains here, mostly overlapping into five categories; two to do with region, three with reference to string, one regarding fabric, one to do with patterning, and one further about lands and foreignness.

3.ii.1.b: FC: Analogical Categories

The analysis above has illustrated both the difficulty of establishing key analogical domains across the full reach of a book-length text, and the difficulties of identifying noise (but see §2.ii.5 above for a justification of the noise-filled approach of this procedure). Nonetheless, it has resulted in some systematic results with regards to the use of analogical domains throughout the full text. These are provisionally identified (in *HT* order) as:

01.02.09.14.01.03 *woven*; 01.03.00 *Physical sensibility*; 01.04.04.01.03 *Viscosity*;
01.05.07.05.09.01 *Hanging/suspension*; 01.06.07.05 *Mutual relation of parts to*

whole; 03.02.07.01.01 A land/country; 03.04.06.14.04 An administrative division of territory; 03.04.06.16 A/the state; 03.07.04.03.08 construction; 03.II.03.01.08.02 Stringed instruments; 03.II.03.01.01.03.01 Tuning/intonation; 03.II.03.02.09.14 Pattern/design; and 03.II.03.04.03.09 Puppetry.

With a relative minimum of qualitative intervention, and based on the principles and processes outlined in §2.ii.3 above, the above domains have been provisionally identified as key analogical domains for this text.

The natural continuation of this investigation would be to look at the lexical activations of this domain, usually through a qualitative corpus concordance. However, one final domain-recognition step is still necessary.

This is because the *HT* conceptual structure, which is necessarily broad and complex, here separates a number of fields which, for the purposes of this thesis, could reasonably be seen as conceptually related. The fine-grained nature of the *HT* therefore can, at times, mask conceptual similarity by placing concepts which could be used concurrently in a text into separate categories. This is not a fault, and is rather due to the precise nature of historical lexicography, which does not aim to reflect the nature of semantic fields in use; the *HT* editors themselves discuss this in the book version's preface (Kay *et al* 2009:xixff). Therefore, when two or more categories have the same referent (consider, for example, the concept of *Fruit*, which was necessarily split in the *HT* into both a product of plants and trees in the natural world, and into a foodstuff for human and animal use), their categorical division is perhaps not valid with regards to how those domains are combined in an authorial-idiolectal text. (By contrast, however, such fine-grained divisions are almost certainly significant for other stylistic uses.) Adapting this precise hierarchy to stylistic use therefore requires one further step, to identify categories in this text which overlap in their use enough to be combined into wider domains.

This can be done using a categorical dispersion plot, which is a straightforward modification of the established lexical dispersion plot visualization technique

used in corpus linguistics. Firstly, the domains are sorted into a rough order based on their meaning, putting categories with similar referents together. This is not a precise task, but does not need to be; its only function is to simplify the visual analysis in the next stage. For *FC*, this could be:

*01.02.09.14.01.03 woven; 01.06.07.05 Mutual relation of parts to whole;
03.07.04.03.08 construction; 03.II.03.04.03.09 Puppetry; 03.II.03.01.08.02 Stringed
instruments; 03.II.03.01.01.03.01 Tuning/intonation; 01.04.04.01.03 Viscosity;
01.05.07.05.09.01 Hanging/suspension; 03.II.03.02.09.14 Pattern/design; 01.03.00
Physical sensibility; 03.02.07.01.01 A land/country; 03.04.06.14.04 An
administrative division of territory; 03.04.06.16 A/the state*

These categories are then examined as if they were lexical items, in a version of the text's Ordered List file. As a reminder from §2.ii.3 above, this is a version of the text with each lemma and the categories it potentially activates in a list, ordered by where it occurs in the text. Taking only the domains and converting it from a list into a long string of category codes gives a version of the text which lists the possible semantic domains in use and where they occur. Looking at this as a lexical dispersion plot (where the lexemes are actually category codes) gives an indication of how much the key domains derived above overlap in terms of their use in the text in question, and suggests how many can be combined.

3.ii.i.b.i: *FC*: Categorical Dispersion Plot

FC's key domain lexical dispersion plot²⁰ is shown in Figure 1 below, where each row represents a semantic domain and a stripe represents a purported activation of the domain at the appropriate place in that text, reading left to right.

²⁰ These and following plots are created using the Python NLTK, which (following tokenization) produces one using the `dispersion_plot` command.

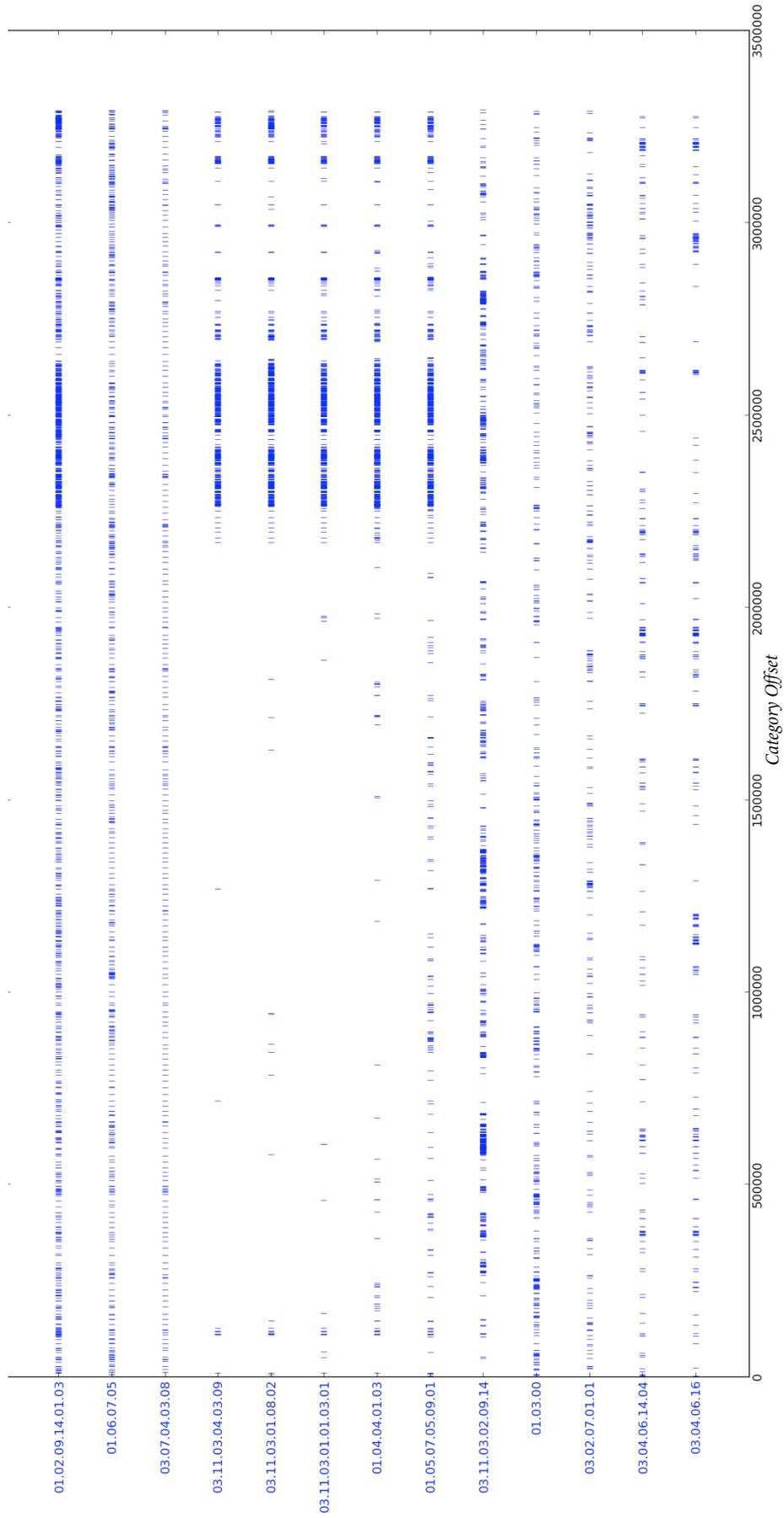


Figure 1: FC Categorical Dispersion Plot

The substantial similarity in distribution between the middle rows of the plot immediately demonstrates the advisability of treating these domains as a group; particularly around the 2.5m offset mark, the five middle domains show overlap to such precision that these most likely represent shared lexemes between those categories. Other domains show limited overlap but substantial differences elsewhere; for example, *01.02.09.14.01.03 woven* has some overlap with rows four to eight, but has a wider use than they do in the earlier part of the text. On the basis of this dispersion plot, eight domain ‘clusters’ can be established for the following final analyses, by merging rows four to eight (the domains of *03.II.03.04.03.09 Puppetry*; *03.II.03.01.08.02 Stringed instruments*; *03.II.03.01.01.03.01 Tuning/intonation*; *01.04.04.01.03 Viscosity*; *01.05.07.05.09.01 Hanging/suspension*) and the final two (*03.04.06.14.04 An administrative division of territory* and *03.04.06.16 A/the state*), and both clusters can be confirmed to have clear semantic overlap by looking at both their category titles and their *HT* entries.

3.ii.I.c: FC: Concordances

The techniques employed here give, after manual sifting for noise, eight key analogical domain clusters, as follows:

1: *01.02.09.14.01.03 woven*

2: *01.06.07.05 Mutual relation of parts to whole*

3: *03.07.04.03.08 construction*

4: *03.II.03.04.03.09 Puppetry*, *03.II.03.01.08.02 Stringed instruments*,
03.II.03.01.01.03.01 Tuning/intonation, *01.04.04.01.03 Viscosity*, *01.05.07.05.09.01 Hanging/suspension*

5: *03.II.03.02.09.14 Pattern/design*

6: *01.03.00 Physical sensibility*

7: 03.02.07.01.01 *A land/country*

8: 03.04.06.14.04 *An administrative division of territory*; 03.04.06.16 *A/the state*

These domains have been checked for accidental ‘noise’ caused by the automatic identification procedure, and have been confirmed to be in use throughout the text. Any clustering of senses around a core has been accounted for by the final clustering above. A further concordance search is not a feasible undertaking within the scope of this thesis (the domains above are realized by 1091 lexical items in the *HT*, which results in 13452 concordance lines of the text); the very nature of the wide-scale analysis done in this part of the thesis makes a concurrent or similarly wide-scale qualitative analysis impractical. However, the following section will give a brief analysis of concordance lines which is intended to be representative of the wide-scale analogical stylistics of this text, even if it cannot be comprehensive. In so doing, it aims to act as a validating supplement to the qualitative analysis presented here rather than a full stylistic analysis of all analogical material in these texts. It also avoids in-depth analyses which pre-empt those in Part II of this thesis, while also supporting those later analyses by assembling some initial contextual indicators of analogical usages.

3.ii.I.c.i: *FC*: Frequent Lexical Items by Key Analogical Domain

To select appropriate lexical items to perform an overview of the concordance lines of the analogical domains identified above, this section identifies the words in each key domain cluster that are most frequent in the *FC* text. These are outlined below, and overlap. Only the most frequent is taken for each domain, apart from those instances where the second-most frequent accounts for more than 50% of the frequency of the first (that is, where the lexical items in the domain do not follow a standard Zipf distribution of natural language, as found in much of corpus linguistics):

01.02.09.04.14.01.03 string, fabric; 01.03.00 feel, sense; 01.04.04.01.03 string;
01.05.07.05.09.01 string; 01.06.07.05 fabric; 03.02.07.01.01 new; 03.04.06.14.04
region; 03.04.06.16 region; 03.07.04.03.08 fabric; 03.II.03.01.01.03.01 string;
03.II.03.01.08.02 string; 03.II.03.02.09.14 figure; 03.II.03.04.03.09 string

These lexical items overlap enough that seven items are all that are required:

string, fabric, feel, sense, new, region, figure

This produces a much more manageable set of concordance lines to overview.²¹ As an incidental point, it is worth noting that none of the items here appear at first to be relevant to the text's literal content, giving support to the domain filtering technique outlined above.

The remainder of this section looks at each item in turn.

3.ii.I.c.I.a: FC: *string*

There are a little over 500 concordance lines for *string* in FC. Browsing the concordance lines shows it is clearly used analogically. This can be seen straight from the beginning of the text, with the following entries:

```
probe which is shaped like a little string And just as a violin string can vibr  
a little string And just as a violin string can vibrate in different patterns ea  
different particle properties A tiny string vibrating in one pattern would have  
rding to the theory such a vibrating string would be what we have traditionally  
ditionally called an electron A tiny string vibrating in a different pattern wou
```

These uses, with the simile-indicators *like* and *just as* (compare the examples in Israel *et al* 2005) alongside the triple use of the hypothetical *would be*, mark the beginning of the analogical form. The text continues to define *string theory*, a particular paradigm within physics and based on this essential analogy. (Many readers of this thesis may well also be aware of this theory, however to be appropriately conservative with regard to the specialist knowledge needed to identify analogical and relevant domains in the methodology outlined above, it

²¹ Concordance lines are again created throughout using Python's inbuilt concordancing function.

was accordingly identified here as analogical even though it is becoming better known outside mathematical and physics circles.) Following this analogical introduction, the text strongly collocates *string* with *theory*, forming a frequent bigram for the next hundred concordance lines. Then further analogical uses begin to appear alongside the bigram:

ifferent vibrational patterns that a **string** can execute It s just like what happ
e those on a violin or cello A cello **string** can vibrate in many different ways a
t musical note In this way one cello **string** can produce a range of different sou
[...]
evel the universe would be akin to a **string** symphony vibrating matter into exist

To the right of the first use of *string* in this excerpt is *just like*, another analogical marker which then precedes a discussion which links stringed musical instruments to the subatomic strings which are the topic of much of the book. These strings, it is worth pointing out here, are postulated to be one-dimensional objects (that is, something with no breadth or depth); it is, and this is an important point, not at all like a physical or musical string. The relationship here is purely a metaphorical one, given that it is currently impossible to view a subatomic string (or prove one exists). Regarding the systematicity of this domain, it follows an interesting pattern, with the introduction of the analogy followed by its entrenchment in the *string theory* bigram, and then interspersed with some uses which relate the analogical use with the original domain of 03.II.03.01.08.02 *Stringed instruments*. At this point, some further collocates emerge; *string vibration* (including *string vibration*, *string vibrations* and *string vibrational patterns*) and *string formulation(s)*. Analogical uses continue with the following cluster:

be one dimensional in which case the **string** s endpoints would be like two beads
two beads sliding on a wire with the **string** itself being like a cord connecting
l in which case the endpoints of the **string** would be very much like two pinballs

Here there are further analogical markers (*would be like*, *being like*, *would be very much like*) and systematic uses of *string* alongside the similar term *wire* (*wire* is likely used to avoid confusion between the musical/physical string and the subatomic string; in which case, further studies here, not taken due to space

limits, may well find that less-frequently-used terms in the key domains above which are synonymous with *string* are likely to have a high analogical use).

This brief overview of a large set of occurrences has aimed to demonstrate the validity of the string/music/viscosity/etc domains being identified above as key analogical domains using only its most frequently-occurring term for reasons of space, and it is clear even from this short analysis that the concept of strings is a significant one. The categories it subsumes have therefore been established quantitatively and checked qualitatively as being key analogical domains with respect to this text.

3.ii.1.c.1.b: FC: *fabric*

The full title of the text under question, *The Fabric of the Cosmos: Space, Time, and the Texture of Reality*, gives an indication of the likely importance of this domain. There are 308 concordance lines for *fabric* in the text, and important analogical uses begin with:

y form the arena of reality the very **fabric** of the cosmos Our entire existence Ñ
[...]
ll cover in this book interweave the **fabric** of space and time in various ways So
ome to wrapping our minds around the **fabric** of the cosmos and touching the true
probability is deeply woven into the **fabric** of quantum reality Whereas human int
r a thick and richly textured cosmic **fabric** Camus declaration notwithstanding de

As a further analogical indicator, the concrete noun *fabric* is here always accompanied by either a premodifier (*cosmic*) or a postmodifying prepositional phrase (*of space and time, of the cosmos, of quantum reality*) to do with physics. It is often accompanied by words from related semantic fields to that of the concrete *fabric* sense – *interweave, woven, richly textured*. Note that the text has had punctuation stripped from it, so that, for example, ‘fabric Camus declaration notwithstanding’ is in the original ‘fabric. Camus’ declaration notwithstanding,’ and so on. The exception to this is the long dash, which is here represented as Ñ, due to limitations in transferring plain text documents between different operating systems.

These uses continue:

untouchable ungraspable unclutchable **fabric** of space is really a something Ñ a s
[...]
ightly probability is woven into the **fabric** of a quantum universe To understand
[...]
eans fundamentally stitched into the **fabric** of classical physics In principle if
[...]
vides another possible shape for the **fabric** of space You can think of this shape
[...]
on a saddle possible shapes for the **fabric** of space We inhabit a single univers
al answer to the shape of the cosmic **fabric** what s abundantly clear is that symm
space and instead imagined that the **fabric** of the cosmos was fixed and unchangi

These continue the pattern above of combining the concrete term with abstract modifiers and other terms from the same concrete semantic neighbourhood.

The analogy has also become established enough to be shifted in its use – there is a reference to the *fabric of classical physics* as opposed to the *fabric of* (for example) *quantum reality*, which modifies the term with regards to the nature of the scientific subject itself rather than the putative focus of the external universe. A second pattern of analogical uses emerges also with the references to *possible shape*, where one feature of fabric – its mutability – is highlighted for later uses:

long an indentation in the spacetime **fabric** caused by the earth s presence or Ñ
[...]
the universe could not be static the **fabric** of space could stretch or it could s
ve had a definite beginning when the **fabric** was maximally compressed and might e
[...]
string theory has revealed that the **fabric** of the cosmos may have many more dim
en dimension entwined in the spatial **fabric** opened new possibilities for experim
[...]
tiny quantum ripples in the spatial **fabric** are also generated by the intense ou
[...]
Gravitational waves Ñ ripples in the **fabric** of spacetime Ñ which were predicted
emergent spacetime in which the very **fabric** of space and time is composed of mor
wn out the microscopic fibers of the **fabric** of the cosmos and unfurled them clea

The analogy is here substantially extended by reference to the variability of fabric's form; it has an *indentation*, it *stretch*[es], it is *compressed*, it has something *entwined* in it, it has *ripples* and *fibres*. The extension here mirrors that which is to be expected from previous metaphor studies, where a new metaphor, once entrenched, can become elaborated on and extended in multiple ways (Lakoff

and Johnson 1980:143ff). *Fabric* is clearly a systematic and wide-scale analogical domain in the text.

3.ii.I.C.I.C: FC: *feel*

There is a substantial amount of interference here from the polysemy of *feel* in terms of emotions, as opposed to its physical sensibility sense. Nonetheless, of the 105 concordance lines for this word there are some consistent uses of it as an analogical technique (an extended list of examples follows):

it s spinning you say and just as we **feel** pressed against the side of a car whe
a concave surface is that Homer will **feel** pressed against the inside of the buc
[...]
s time if you start spinning will you **feel** it Will your arms and legs feel pulle
l you feel it Will your arms and legs **feel** pulled outward Our experiences in day
from not spinning a state in which we **feel** nothing to spinning we feel the differ
[...]
meaning It s not just that you won t **feel** anything if you spin it s more basic
[...]
tion of accelerated motion If you can **feel** spinning motion in a universe with me
in an empty universe you would still **feel** the sensation of spinning Mach offere
[...]
Scratchy and Apu on the train do not **feel** any motion From their perspective it
[...]
upward Nevertheless the forces you d **feel** struck Einstein as very familiar As y
eyes during the descent Barney would **feel** exactly what he would if he were floa
will change significantly Barney will **feel** pressed to the capsule s floor just a
[...]
t local The Red and the Blue To get a **feel** for the kind of nonlocality emerging
[...]
properties the system has To get a **feel** for the ideas imagine unbinding a cop
[...]
line in a 100 meter race And so we do **feel** the Higgs ocean The forces we all exe

These give a sense of the somewhat limited way in which *feel* and the rest of *01.03.00 Physical sensibility* can be seen as analogical, usually within the context of extended stories of physical actions used to give context to the discussions of gravity, motion and spinning. The use of either the second-person pronoun or a range of lighthearted names (taken, in this text, almost exclusively from the cartoon *The Simpsons*) here acts as a distinguishing feature of these examples. There are also some uses (for example, the second and third from last in the list above) of the entrenched metaphor UNDERSTANDING IS GRASPING. This domain is clearly an analogical one, but it is somewhat different from the others above; instead of relating two different domains, it analogizes using embodiment

sensations, which differ somewhat with regards to their cognitive status as opposed to semantic domains (see Gibbs 2006b, or Part IV of Evans and Pourcel 2009).

3.ii.I.C.I.d: *FC: sense*

This differentiation is mirrored in the 98 occurrences of *sense* in the text. Excluding those due to the alternate meaning of *sense* as *meaning*, a large number are embodied, generally with reference to intuition:

suggested revisions to our everyday **sense** of reality that are as dramatic as m
[...]
ues unsatisfied Beyond the intuitive **sense** that no perspective should be more r
[...]
formly in concert with our intuitive **sense** of time comparison with other clocks
[...]
physics that supports this intuitive **sense** that time flows In fact a reframing
ime As a prime example the intuitive **sense** of time s flow can be vividly portra
e right now As I type these words my **sense** of what exists right now my sense of
my sense of what exists right now my **sense** of reality amounts to a list of all
ntal freeze frame image captures our **sense** of reality our intuitive sense of wh
s our sense of reality our intuitive **sense** of what s out there it consists of e
[...]
of the very things that dominate our **sense** of reality Again there is no questio

Leaving aside the tangential issue of what it means to *sense reality* as opposed to *sense* in its normal meaning, the embodiment analogies are clear in this section, supporting the inclusion of *01.03.00 Physical Sensibility* as a key domain while encouraging its separation in later analyses due to its embodied, as opposed to domain-based, analogical status.

3.ii.I.C.I.e: *FC: new*

This item appears problematic, and the category itself (*03.02.07.01.01 A land/country*) was only provisionally labelled as analogical above, and then only due to its lexical spread. The selection of *new* as a lexical item is similarly problematic, and analysis of the 172 occurrences of this word show it is used solely in its core sense of novelty. The ranked list of items instantiating this domain is:

<i>HT</i> Category	Lemma	Frequency
03.02.07.01.01	new	172
03.02.07.01.01	great	44
03.02.07.01.01	land	23
03.02.07.01.01	mother	23
03.02.07.01.01	home	15
03.02.07.01.01	exotic	15
03.02.07.01.01	interior	14
03.02.07.01.01	limitation	11
03.02.07.01.01	settled	8
03.02.07.01.01	seat	3
03.02.07.01.01	stranger	2

Some of these are certainly noise, and analysis of each alternative does not reveal many analogical uses (examples of common uses which do not have analogical content include *great time, great uncle, Great Wall of China, great British physicist; the particle is likely to land where the wavefunction value is large; the credit for my mother's groundbreaking paper, in due course your mother gave birth to you; closer to home, this place we call home, from my home in New York; exotic results, exotic particles; the black hole's interior* and so on), and the few which are analogical are by no means systematic enough to draw the attention of the present analysis. Without the presence of any substantial or systematic analogical uses of this domain in the text, it is removed from the list of key domains, and the importance of looking qualitatively at the actual text following any quantitative analysis is thus emphasized.

3.ii.I.c.I.f: FC: region

This item occurs 109 times in the text. It generally collocates with *of space* in a non-analogical way, and has very few analogical uses. Looking at the list of key domains, this term represents *03.04.06.14.04 An administrative division of territory* and *03.04.06.16 A/the state*, which (as with *new*) was identified with some uncertainty as key above. This, along with its lack of analogical uses in its main lexical realizations, is enough to disqualify it as analogical and thus remove the full cluster.

3.ii.I.c.I.g: FC: *figure*

The concordance of the 387 occurrences of *figure* is heavily noise-filled due to the captions and cross references throughout the text (see *Figure 15*, etc).

Removing these, and then removing the *reckoning* meaning of the bigram *figure out*, gives no analogical senses of this word. The subsequent lexical realizations of the domain 03.II.03.02.09.I4 *Pattern/design* are *wave* with 144 occurrences and *pattern* with 125, and so are analysed here instead.

3.ii.I.c.I.h: FC: *wave*

From the text's concordances (as well as from general knowledge), it is clear that *wave* has both literal and analogical senses in physics (*light/electromagnetic wave*, for example, as well as a particular regular curved to-and-fro pattern), and that the text attempts to use one to explain another:

```
ter toward an oncoming water wave the wave approaches you more quickly if you sw  
ore quickly if you swim away from the wave it approaches you more slowly Similar  
[...]  
lusion light unlike any other kind of wave ever encountered does not need a medi  
[...]  
I see direct evidence that light is a wave If you ve never done this it s worth  
ly our understanding to light A water wave disturbs the flat surface of a lake b  
[...]  
s means picture a snapshot of a water wave that shows regions of high intensify  
y the greater the potential the water wave has for exerting force on nearby ship  
[...]  
the analog of a gently rolling ocean wave shown in Figure 4 6 Since the peaks a
```

Analogical indicators here are more lexical than were found previously, and include the markers *our understanding*, *picture* (as a verb) and *analog*. There is a further question here as to whether the analogy is with regards to the periodic pattern or to the water wave; in fact, the domain 01.01.05.00 *Water* has a log-likelihood ratio value of 442 in this text, which is enough to make it significant without pushing it into the top 100 domains analysed above. An argument can easily be made that this is in fact a three-domain analogy linking water waves and electromagnetic waves through the periodic pattern; this idea will be

returned to in a later chapter. For now, it will be assumed that were *01.01.05.00 Water* to be analysed here, it would be merged into the same analogical cluster of domains as the present *03.II.03.02.09.14 Pattern/design* domain, on the basis they are similar enough to warrant clustering. Given this, the question of whether the waves here are water waves analogically applied to electromagnetic waves or wave patterns applied as a link between similar liquid and electromagnetic phenomena is moot.

3.ii.I.c.I.i: *FC: pattern*

Pattern collocates here quite often with *interference* and *vibrational* (referring to string vibration; see above). It has a few occurrences which lend support to the wave-as-pattern argument above:

ey yield the undulating probability **pattern** of wave interference And so unlike
[...]
move in concert creating a wavelike **pattern** in their overall macroscopic motion
[...]
ions generate follows a distinctive **pattern** of steeply rising then gradually fa

This is in addition to forming some links between the pattern analogy and other key domains above:

any different ways and we hear each **pattern** as a different musical note In this
[...]
s a string vibrating in yet another **pattern** And of prime importance what Schwar
[...]
ed over a muted din or an elaborate **pattern** embroidered on a plain piece of mat

Finally, it has a limited few of its own analogical uses, signalled lexically, for example:

you imagine drawing a checkerboard **pattern** on the event horizon of a black hol

These together, ignoring the noise of *figure*, help confirm this domain as a key analogical one in this text.

3.ii.1.d: FC: Summary

The analysis above aimed to qualitatively test and verify the empirical analysis already presented. It succeeded, in the main, verifying six of the eight key analogical clusters identified above (§3.ii.1.b.1), and eliminating two from consideration. This is a more than acceptable ratio, and both confirms in the main the identifications above while providing some additional evidence to strengthen (and reduce) the list of domains. Some further work as to the role and discourse presentation of analogous material was also usefully performed in advance of later analyses. Following consideration of the remainder of the corpus, these results will be used in the following section to assist identification of sections for conceptual blending analyses in later chapters of the thesis.

One final note: the *stringed* analogy, known now as *string theory*, has already been analysed in this section. It will therefore be treated in the remainder of the corpus as a relevant domain, for the purposes of the present part of this thesis. Again, popular science readers may have already been aware of the existence and analogical use of string theory, but will note that as it requires somewhat specialist knowledge to understand it as a feature of texts in this area, identifying these domains as relevant would not be appropriate under the analytical guidelines outlined above.

3.II.2: PW: PARALLEL WORLDS

Note again that the length of the above *FC* analysis was heavily detailed in order to illustrate the process; the identification of key, relevant and noise-filled domains is here abbreviated.

3.ii.2.a: PW: Key Domains

The eight key domains of *PW* are:

03.08.04.06.02 Hand gesture/wave

*03.II.03.01.01.03.01 Tuning/intonation **

*03.II.03.01.08.02 Stringed instruments **

*03.II.03.04.03.09 Puppetry **

01.05.07.03.06.01 Series of curves

01.05.07.05.04.02 Making holes/becoming holed

01.05.08.02.02.01 Oscillation

03.01.03.02 Civilization

Of these, the three starred domains of *03.II.03.x* ‘string-ness’ are set aside in the below analysis, as they have been dealt with above, leaving six key analogical domains for investigation here. Note that the final four were only tentatively identified as key based on the criteria in §2.ii.3 above.

The relevant domains for this text came from *HT* sections *01.01.10.x The Universe*, *01.04.07.x Physics*, *01.04.08.x Light*, *01.05.01.x Existence*, *01.06.06.x Quantity*, *02.01.x Mental capacity*, *03.02.07 Inhabited place* (cf *FC* above) and *03.04.06.x Government*, as well as the more specific domains *Darkness*, *Instrument for distant vision*, *Atomic chemistry*, *Solemnity*, *University*, *Base metals* and *Puzzle*. Noise was primarily due to the polysemy of *star*, *vacuum* (cf *01.02.II.01.09 Vacuum-cleaning*), *light*, the colour *black*, *hole* (cf *03.II.04.13.16.12.01 Golf-course*) and the expected *02.08.02.06 Indo-Hittite*.

3.ii.2.b: PW: Analogical Categories

3.ii.2.b.I: PW: Categorical Dispersion Plot

The categorical dispersion plot for the above domains is shown in Figure 2.

The similarity between the first, second and fourth rows of the plot identify those as best grouped; they are *03.08.04.06.02 Hand gesture/wave*, *01.05.07.03.06.01 Series of curves* and *01.05.08.02.02.01 Oscillation*. This gives three key domain clusters for this text.

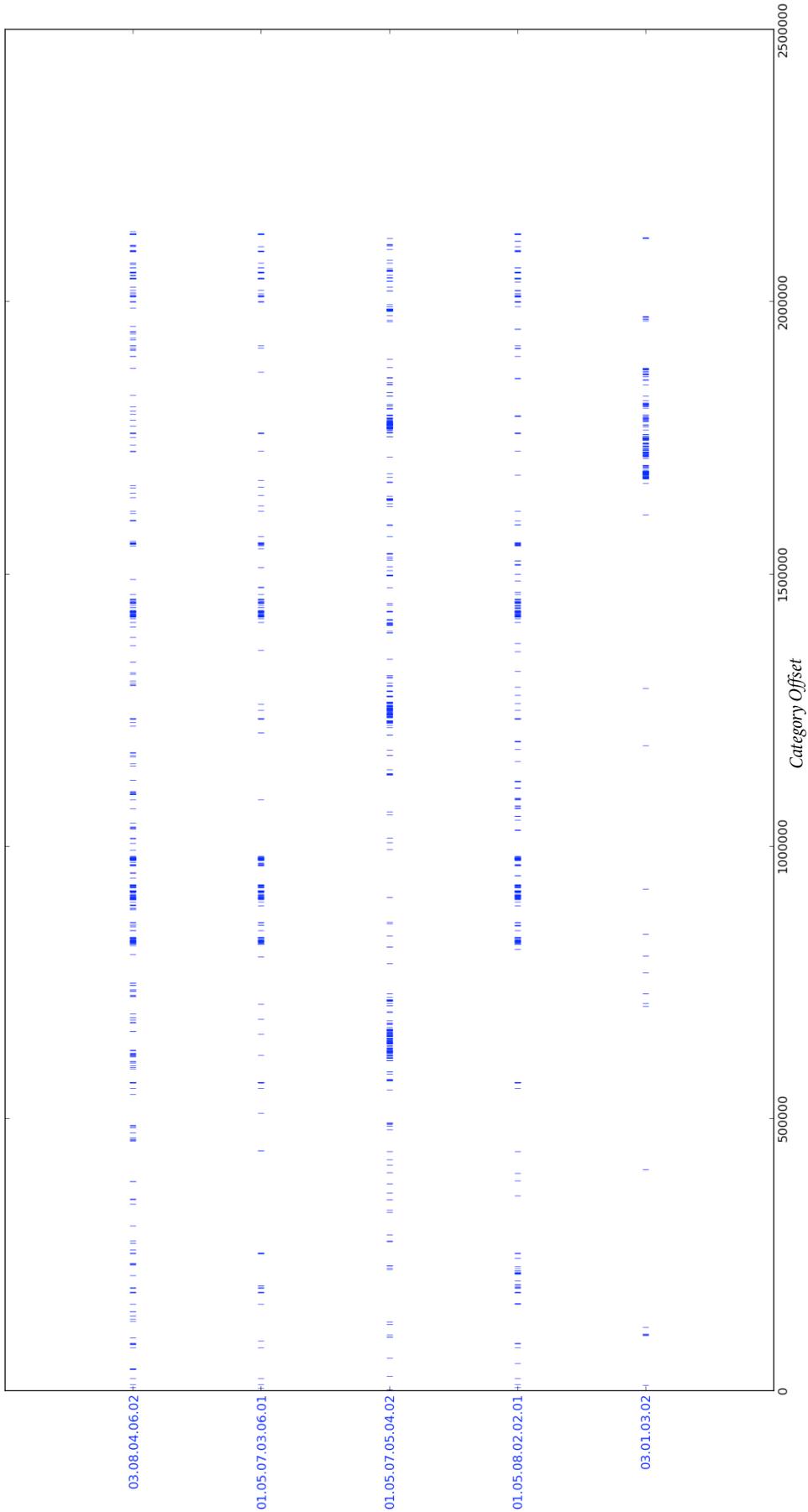


Figure 2: PW Categorical Dispersion Plot

3.ii.2.c: *PW*: Concordances

3.ii.2.c.1: *PW*: Frequent Lexical Items by Key Analogical Domain

Again, here the most frequent lexical item is taken for each domain, apart from those instances where the second-most frequent accounts for more than 50% of the frequency of the first. This gives:

01.05.07.03.06.01 wave; 01.05.07.05.04.02 hole; 01.05.08.02.02.01 wave;
03.01.03.02 civilization; 03.08.04.06.02 wave, point

Removing repeated words, this gives *wave, hole, civilization* and *point*.

3.ii.2.c.1.a: *PW*: *wave*

While *wave* has been looked at above in *FC* in the sense of a pattern, it is here key in the sense of a hand gesture. This is, again, a reflection on the lexical makeup of similar domains. Unlike *FC*, though, *wave* here collocates with *function* heavily, and is used purely in a relevant, non-analogical technical sense. It is therefore removed from the list of key domains for this text.

3.ii.2.c.1.b: *PW*: *hole*

Hole's 212 occurrences are in the main linked to the bigram *black hole*, which accounts for the vast majority of these uses. It is difficult to distinguish explanatory analogies from established terms which are based on analogy (like the *string* terminology above, only much more pronounced in this case); the question which requires an answer is this: is *black hole* being used in this text in such a way that it is analogical, or in a standard literal scientific way?

A number of indicators are here suggested based on the analogical material thus far analysed. Firstly, the presence or absence in the concordance lines of markers of analogy, as introduced in the *FC* analysis above, can indicate analogical usage, for obvious reasons. Secondly, the presence of extended analogical terms can do the same, rather like the *fabric* and *string* examples

above being extended into *wire, thread, loop, patterns*, etc. Thirdly, the use of the same term simultaneously in multiple domains (like the *wave* example above) is significant here.

Examples can be found of each of these. Firstly, there are markers of analogy such as second person pronouns:

ld radius As you approached the black **hole** in a rocket ship you would see light point you can see surrounding a black **hole** there is a magic sphere at the Schwarze parallel universes for a Kerr black **hole** but you can not return once you enter No matter what you throw into a black **hole** you lose all its information This goe

There are also references to the appropriateness of the term with reference to reality:

ekenstein and Stephen Hawking a black **hole** is not perfectly black because it slo

See also the *punch a hole* example below. Secondly, with reference to extended analogical terms, there are a few to be found:

Rosen bridge At the center of a black **hole** there is a throat that connects s radiation that tunnels out of a black **hole** There is a tiny but calculable probab [...] r the topology of space by punching a **hole** and leaping into hyperspace It simply

But these are not systematic. Note that the replacement of punctuation with spaces has resulted in the loss of the quote marks around *throat*, which are found in the original.

Thirdly, there are instances of *hole* being used literally with reference to black holes analogically. For example, the first time the concept is introduced it is referred to as a *hole in space*, then later the same object is called a *black hole*:

s a powerful atom smasher to create a **hole** in space that connects our universe t ld radius As you approached the black **hole** in a rocket ship you would see light

In addition to a further analogical digression comparing what would happen if a donkey fell into a real hole and a black hole:

exotic object in the universe a black **hole** which I discuss in chapter 5 68 BIRD k of a donkey that falls into a large **hole** in the ground We have to add energy t

monkey in order to pull him out of the **hole** Once he is out and he is standing on
have had negative energy while in the **hole** Similarly it takes energy to pull a p

In this final example, *similarly* also acts as a marker of analogy.

These factors, in combination with the analogical nature of the term originally,
is enough to support it being considered an analogical domain, but with some
caution needed when applying any further analyses below or in Chapter 4.

3.ii.2.c.I.c: *PW: civilization*

The concordance lines here for *civilization* suggest this domain be considered as
literal; while it was not immediately relevant from the paratext of the book
(but it is to a reader; although remember above that the methodology does not
use prior knowledge of the text), the text spends some time discussing
encountering alien civilizations:

another universe Can an advanced **civilization** perhaps millions to billions of y
otent of an advanced intelligent **civilization** to the molecular level and inject
[...]
civilizations in outer space Each **civilization** type emits a characteristic form
ed and cataloged Even an advanced **civilization** that tries to conceal its presenc
[...]
iverse The concept of an advanced **civilization** creating an escape hatch to anoth
[...]
cessary to regenerate an advanced **civilization** the resources to do this raw mate

This can also be seen by looking at the line on the dispersion plot for
03.01.03.02 *Civilization*, which clusters around one particular area. This is the
chapter on alien civilizations. An argument is then made for not accepting
domains as key over the whole text whose entries on dispersion plots form
single clusters.

3.ii.2.c.I.d: *PW: point*

The final term, *point*, is similarly used literally throughout all its occurrences,
and the domain is thus removed from analogical status.

3.ii.2.d: *PW*: Summary

Of the four analogical domains postulated above, only one, *hole*, remains as a wide-scale analogical one, and this is a tentative analysis. This is a useful result; by comparison with *FC*, it can be hypothesized here that *PW* has no equivalent unifying analogies which operate across the whole text systematically. This could be for a variety of reasons, but one which is immediately obvious is that *PW* has a different scope of genre than *FC*: it is a tour of different ideas in modern cosmology, rather than a single coherent monograph on a sole topic. A local analysis, as is undertaken later in this thesis, can establish if analogy is not used in this text, in contrast to the claims of the Introduction above, but it is an unlikely outcome. In which case, domains of local semantic relevance are more significant than global ones. This idea will be returned to in the analysis of *WP* below.

3.II.3: *MP: THE MUSIC OF THE PRIMES*

3.ii.3.a: *MP*: Key Domains

MP has a high number of relevant domains and noise, giving only two putative key analogical domains. These are *01.05.07.01.01 Distance/farness* and *01.05.07.06 Direction*.

The relevant domains for this text came from *HT* sections *01.06.04.x Mathematics* and *02.01.x Mental capacity*, as well as the more specific domains *Nation*, *Infiniteness*, *Logical syllogism*, *Wishing*, *Determination*, *Persistence*, *City*, and *A/the state*. Noise was primarily due to the polysemy of *prime* (cf *03.II.04.I3.03.07 Fencing*, *03.10.II.42.03.04 Boiler*, *03.07.03.06.03 6am*, etc), *number*, *line*, and *proof*.

3.ii.3.b: *MP*: Analogical Categories

3.ii.3.b.1: *MP*: Categorical Dispersion Plot

The *MP* dispersion plot for these key domains is Figure 3 below. There are no groups to be found, although it is noticeable that the lack of other purportedly

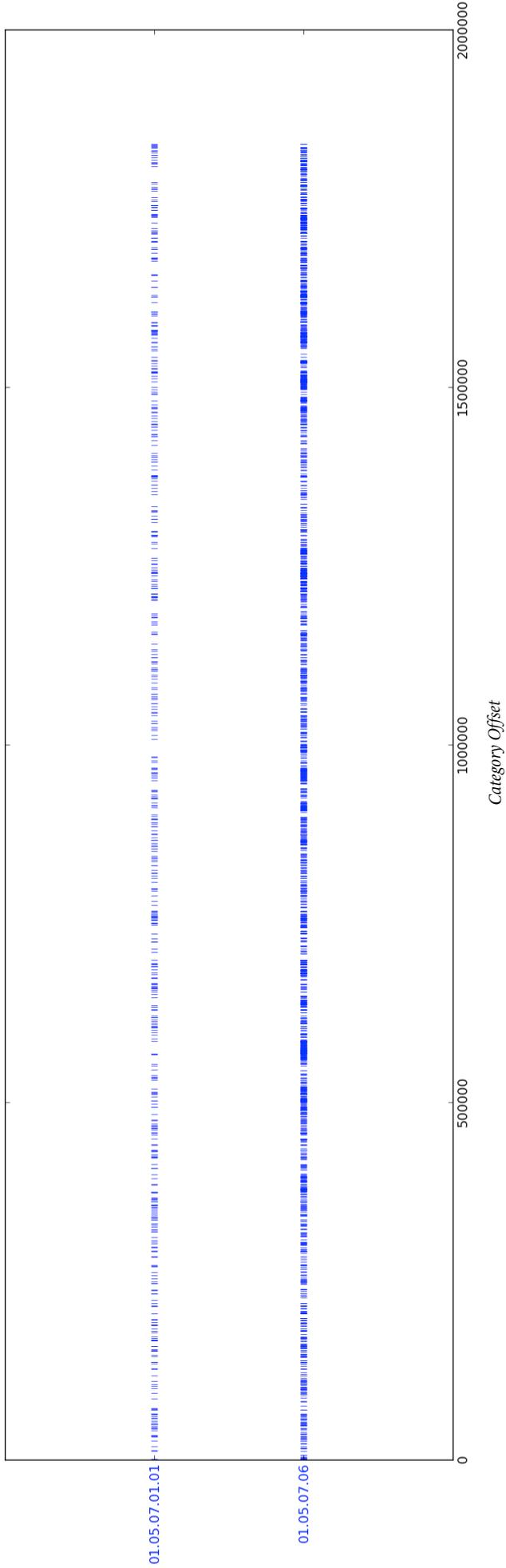


Figure 3: MP Categorical Dispersion Plot

analogical domains in this text appears to have resulted in a much wider spread of *01.05.07.06 Direction* than that of key analogical domains in other texts.

3.ii.3.c: MP: Concordances

3.ii.3.c.I: MP: Frequent Lexical Items by Key Analogical Domain

The most frequent lexical items in these two categories are *way*, *far*, *line*, *level* and *point* (the relatively large number of words here is because *01.05.07.06 Direction* has a wide lexical spread before the 50% cut-off point).

3.ii.3.c.I.a: MP: *way*

While *way* has a methodological meaning that is often used in science, the 198 concordance lines of this word in *MP* show a significant analogical use. They recur with regards to an extended analogy of a landscape or topography:

```
blem As mathematicians navigate their way across the mathematical terrain it is
[...]
0 bc they had evolved a very physical way of understanding what it is amongst a
[...]
thor Not only do they finally see the way to the peak but also they understand
new mathematics encountered along the way The problems allow for exploration of
ling of exhilaration at discovering a way to reach the summit of some distant p
[...]
ily rigid geometry There was only one way that the landscape could be expanded
[...]
r case mathematicians finally found a way to avoid having to cross the summit o
[...]
dy and Littlewood were fighting their way across Riemann s strange landscape so
[...]
height of the Riemann landscape This way he might be able to find a point at s
[...]
important breakthrough He had found a way to show that points at sea level in a
othesis for primes but he had found a way to show that points at sea level in r
```

There are some important methodological points to be noted here before examining the concordance extracts stylistically. Individual items here may be easily mistaken, in the absence of much longer concordance lines, as literal references (contextual information as to where events took place, perhaps, or actual landscape analysis). It is only when viewed *en masse* (and taken into account with items later in this section) that its relevance becomes clear, even though slightly longer concordance lines were needed in some cases to confirm

their use. In this way, the importance of concordance analysis in corpora, and in examining where possible the full range of a text is illustrated (see Sinclair 1991).

The core of the widespread analogy here is found in some of the lexical collocates and phrases listed above – *terrain, a very physical way of understanding, peak, encountering on the way, summit, landscape, sea level*, and so on, referring to a mathematical formula. The landscape analogy, spanning the full range of the concordance and the text (see the dispersion plots above), is also described using the lexical items below, and so a full description is reserved until the end of this section.

3.ii.3.c.I.b: *MP: far*

The 91 occurrences of *far* also contribute to the landscape analogy. They include such uses as (a lengthy list follows, to illustrate the breadth of the analogy):

s could see no rule that told him how far to jump to find the next prime Were m
[...]
ghting of the possible existence of a far off mountain like Fermat s Last Theor
[...]
primes never run dry Or perhaps this far off peak is just a mathematical mirag
[...]
further you count Gauss had seen this far off mountain peak but it was left to
[...]
os in his imaginary map Remarkably so far his analysis had worked without him h
[...]
formula which were determined by how far east each of Riemann s zeros would be
uld prove that there were no zeros as far east as the border through 1 Finally
[...]
n to obsess him but the others ranged far and wide over the mathematical landsc
[...]
Littlewood could not even predict how far we would need to go before we could o
[...]
be seen until one had travelled very far north in the landscape As Littlewood
[...]
al test By the time you had gone this far north if the graph still did not cros
[...]
s finally powerful enough to navigate far enough north in Riemann s zeta landsc
[...]
ician Hugh Montgomery about the zeros far up Riemann s ley line Odlyzko recogni
[...]
ng the largest percentage of zeros so far to be found on Riemann s line Fry emb
[...]
up in a straight line He couldn t see far enough across his landscape to tell w
are patterns to be discovered in the far reaches of the numbers Maybe the prim

These are substantial analogical uses here of this item, contributing heavily to the landscape analogy, and identifying it with a graph for the first time thus far. Again, a detailed analysis is presented at the summary at the end of this section; for now, this domain is clearly key.

3.ii.3.c.I.c: *MP: line*

Line occurs 181 times in the text, and collocates heavily with *number* to form *number line*. While this could be argued as an entrenched metaphor (used in arithmetic since René Descartes (1596-1650) and in education for at least the past few centuries; see further Mazur 2004), there is a very strong case to be made that in the light of the above landscape activations that the number line can be considered an active analogy in this text. As with §3.ii.I.c.I.d above, all three proposed indicators of analogical usage (explicit markers, extended analogical items, simultaneous lexical use in multiple domains) can be found in the number line examples. In particular, its extended usage in non-entrenched contexts is substantial. Significant occurrences, including those which establish analogical relevance in addition to those which aid understanding of the landscape analogy follow:

n what mathematicians call the number **line** The square root of 2 for example is a
 umbers took their place on the number **line** stretching out to the left of zero Th
 [...]
 ordinary numbers as lying on a number **line** a line running east west on which eac
 [...]
 e Greeks But there was no room on the **line** for this new imaginary number the squ
 What if one unit north of the number **line** were used to represent i All the new
 [...]
 ning up as if along some mystical ley **line** running through the landscape Riemann
 [...]
 l the same Riemann knew that this ley **line** had an important significance He alre
 [...]
 that Nature would also have used this **line** of symmetry to order the zeros What i
 [...]
 hem must actually lie on the critical **line** Such are the mysteries of mathematics
 [...]
 evel in the zeta landscape are on the **line** through Without making this assumptio
 [...]
 andscape were indeed on Riemann s ley **line** But Turing had seen an even more soph
 t at sea level off Riemann s critical **line** and prove the Riemann Hypothesis fals
 [...]
 hills and valleys along Riemann s ley **line** Zagier s graph represented a new pers
 [...]
 zeta landscape all lay on a straight **line** In the year following his discoveries

[...]

e there was hope that they would also **line** up in the landscape of the primes Not

These extracts are only a sample of the possible examples; of the 181 uses of *line* in this text, 170 are analogically relevant (the remaining 11 relevant uses are generally those where *line* is a component of a phrasal verb such as *line up*). The use of the term *i* above is not an error, but rather a mathematical symbol referring to the square root of -1. A few facts are worth noting here. Firstly, the bigram *ley line* is fairly frequent, which instantiates a further domain of esoteric knowledge alongside the directionality one above. Secondly, the *landscape* term is often premodified by *zeta*, giving a further bigram which is frequent enough in the text to be used here as the title for this particular analogy.

3.ii.3.c.i.d: *MP: level*

The 72 uses of *level* in this text separate into 64 occurrences of the bigram *sea level*, and 8 further uses of *level* individually. Again, given the zeta landscape analogy already described, it should come as no surprise that *sea level* is here analogical:

a smooth plane 1 unit high above sea **level** If Riemann turned round and started
ints where the landscape fell to sea **level** could reconstruct everything about t
[...]
ordinates of all these points at sea **level** would give him enough information to
all the hills and valleys above sea **level** Riemann did not forget where all his
prime numbers and the points at sea **level** in the zeta landscape was about as d
where the zeta landscape was at sea **level** he could get rid of these errors and
e The farther north the point at sea **level** the faster the wave corresponding to
zeta landscape and the points at sea **level** in this landscape are the key to unl
whereby each of these points at sea **level** produces a wave like a musical note
[...]
ted from the zeros the points at sea **level** in the landscape were like the sound
[...]
the coordinates of the points at sea **level** in the landscape corresponding to th
[...]
has calculated so many points at sea **level** in Riemann s treasure map believes w

There are many points where *sea level* collocates within ten words of *landscape*, and there are some brief further musical analogies not yet encountered.

3.ii.3.c.1.e: *MP: point*

Finally, *point* is more noise-filled than the previous lexical items but still has some pertinent uses here. It particularly collocates with the existing analogical terms above:

e for the wave The farther north the **point** at sea level the faster the wave cor
nfinite number of waves one for each **point** at sea level he encountered as he he
ld that sat at sea level Out of each **point** he had created a wave like a note fr
[...]
ich described lines and points and a **point** could be changed into numbers descri
[...]
e and the notes he created from each **point** at sea level is music utterly unique
[...]
ed into music In this landscape each **point** at sea level sounded a note The sear

Here, the uses of musical analogies are continued from *level* above, and *point* collocates with *level*, *line*, *landscape*, *farther*, and others. This illustrates that *point*, while not as extensively used as the other items, is nonetheless a contributor to the extended zeta landscape analogy, which is looked at below.

3.ii.3.d: *MP: Summary*

The extended analogy here has been represented in all of the key analogical domains identified, and as such is the most widely-coherent analogy thus far. It has been termed here the ‘zeta landscape’ analogy, after the most specific reference to it which was found in the concordance lines above. The text’s introduction to this idea is:

Gauss's two-dimensional map of imaginary numbers²² charts the numbers that we shall feed into the zeta function. The north-south axis keeps track of how many steps we take in the imaginary direction, whilst the east-west axis charts the real numbers. We can lay this map out flat on a table. What we want to do is to create a physical landscape situated in the space above this map. The shadow of the zeta function will then turn into a physical object whose peaks and valleys we can explore.

²² The Gaussian reference here is to the space mapped out by the two number lines, where one stretches west-east for real numbers (1, 2, 3, etc) but north-south for imaginary numbers (a number which gives a negative result when squared; this cannot be a number as the term is commonly understood, thus the term *imaginary*). They meet at 0. See further Rozenfeld 1988.

Here du Sautoy introduces his core analogy, and proceeds throughout the book to refer to and elaborate on this. Using only terms found in the concordance above, the *map* refers to *treasure* and then becomes a *physical terrain*, with *exploration* taking place, and with *mountains*, *peaks* and *summits*, arranged *north* and *east* around a *border*; the *landscape is strange* and rises from *sea level*; it is *traversed* by a particular *straight critical line*, sometimes called a *ley line*. Looking at its chronological development, the analogy begins as a map or abstraction arising from the abstract visualization of the number line, and is then transformed from a graph to a landscape following the contours of the graph. This landscape is progressively elaborated and frequently referred to in the text (see the dispersion plot above).²³

3.II.4: WP: WARPED PASSAGES

3.ii.4.a: WP: Key Domains

WP is here unusual in that once its eight *string* related key domains were removed, only one domain was identified as being possibly analogically key: 03.07.02.01.10 *Hierarchy*. This was, moreover, established by only one lexical item: *hierarchy* itself.

The categorical dispersion plot for this category is in Figure 4. The category appears to cluster around two points, one around the 1.5m mark and a further around the 2.25m mark. This disqualifies it as a systematic and recurring analogy, and a concordance search for *hierarchy* reveals it refers to the *hierarchy problem* (a particle physics problem to do with the scale of universal constants).

This presents a difficulty for the present analysis procedure, as it means that the hundred most key domains in WP are non-analogical. This could be due to one of three different hypotheses:

²³ There also, from the above extract, appears to be a local shadow-related analogy which the larger one interacts with at this introductory stage.

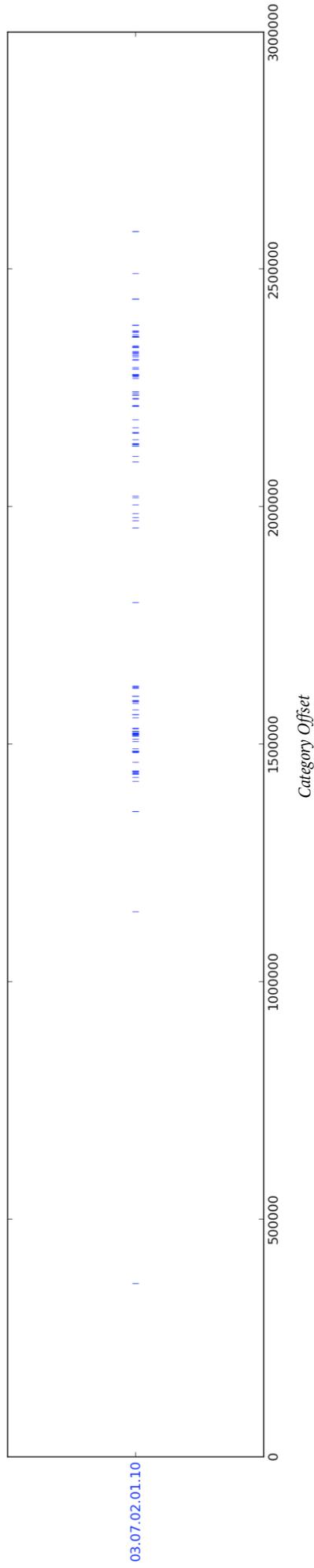


Figure 4: WP Categorical Dispersion Plot

1. This text is non-analogical.
2. The threshold of 100 domains is too low to discover systematic analogical domains in a text with a substantial amount of noise.
3. The analysis here is insufficient for a text which has multiple varying analogies but no systematic text-wide analogical domain.

These three cover the range of possibilities as they address firstly the nature of analogy (it is either not there, as in 1, or it is there but undetectable, as in 2 and 3) and the range of undetectability (due to either the – arbitrary – methodological limit of 100 as in 2, or due to the nature of the text as in 3). If none of these hypotheses are correct then a further hypothesis would be that the methodology outlined in this thesis is unusable, which is an unlikely outcome given the analyses above and the comparison (that is, confirmation) analyses below. Each of the three numbered hypotheses above are examined in the following sections.

3.ii.4.b.i: WP: Non-Analogical Text?

There are a number of measures which might be taken to check if a text is purely non-analogical. One way is a concordance check on terms in the lexical field of analogy. The word *analogy* itself appears 33 times, with such concordance lines as:

they may be easier to understand by **analogy** In 1884 to explain the notion of ex
such a tiny size So to complete our **analogy** suppose that something much bigger
probably the simplest is through an **analogy** Imagine that you throw a ball up in
cessive high frequency radiation An **analogy** might help elucidate Planck s logic
orbits That is why I could make the **analogy** with different floors of a tall bui
late 1800s because of a tangential **analogy** to railroad gauges that tell you th
ve you the general idea by using an **analogy** Suppose you have a shirt making mac
ing to one another as well A better **analogy** for the contributions due to virtua
are present in equal numbers As an **analogy** imagine an equal number of differen
this phenomenon with the help of an **analogy** Suppose that you prepare six table
this with the aid of the sprinkler **analogy** from Chapter 20 A flat extra dimens
receive Because I will use the duck **analogy** again in the next chapter I ll expl
y in a localized region A real life **analogy** of this sort would be the Colorado

These examples, selected from a longer list, show that there are clear concrete analogies in the text (featuring *balls, floors of a building, railroad gauges, shirts, tables, sprinklers* and *ducks*); and these examples are only those in the immediate context of an explicit lexical item such as *analogy*. By comparison, the *MP* text shown above to have extensive analogical reference to a landscape only has four occurrences of *analogy*, two of which could be considered noise, and the *FC* text has fifteen. It is clear the text does use analogy. This investigation then continues by considering two opposing possibilities for the lack of key domains identified in this text; have these key analogies simply not been found because of the limited number of key domains analysed, or, approaching the problem from the opposite angle, are there instead no overarching key analogies but rather a number of smaller, local analogies?

3.ii.4.b.2: *WP*: Threshold Too Low?

The 100-domain threshold established in the methodology above was a pragmatic one; with seven texts, this gave 700 domains to be investigated, searched for, and examined with regards to its lexical spread and word instantiations. This was felt to be reasonable given the scope of this thesis and the scale of the domains in use. The lack of analogical domains in the first 100 selected by the procedure outlined above does, however, raise some questions as to whether this limit is too small to discover systematic analogical domains in use in this text. The sample size for *WP* was therefore doubled, and the following hundred domains analysed as the first were.

There were, however, no analogical domains in this set either; the extra hundred contained 28 further relevant domains (including such categories as *Electromagnetic radiation, Relativity, Trial/experiment, Experience, Science*), and 72 further sets of noise. The possibility remains that there are further analogical domains further down the log-likelihood ranking, although this argument quickly becomes recursive and self-justifying if the threshold is lowered progressively lower until a result is found. The pragmatic limit above sufficed

for the comparison texts and three of the four core texts; even when doubled, it produced no analogical results. It is difficult to justify increasing the limit repeatedly,²⁴ and so the next stage is to work from a related assumption: that the ranking will not produce analogical domains as they are too varied to produce a keyness result.

3.ii.4.b.3: *WP*: Multiple But Not Text-Wide?

This final possibility is also the one most supported by the short and unrepresentative *analogy* concordance test above; it is what remains once the two hypotheses above were eliminated. To be clear, any text with widely-spread numbers of analogical domains, which are each relatively under-activated lexically by comparison with those texts above with overarching analogical material, will not be recognized as key with respect to the log-likelihood ratio of a wide-scale reference corpus. This is an inevitable consequence; the reference corpus needs to be such that any small-scale domains of this type will not be recognized as significant with reference to the full text, even though their removal causes some problems for this particular text. Therefore, *WP* will count as a text with no systematic overarching analogical domains, but does need to be established as a text with analogous material.

To capture small-scale analogies, a different set of procedures need to be adopted. We can run a concordance analysis like the *analogy* test above, but using a more frequent marker of analogy such as *like*. This gives 252 occurrences, the majority in the prepositional/conjunctive-analogical sense (that is, not in a noun, adjective and adverb pattern, and not those conjunction uses with a perception verb preceding *like*, which generally indicates the *as though* sense of *like* rather than the analogical one, which tends in these samples to be followed by a copula).

²⁴ See further Karl Popper, who, when discussing the philosophy of science and falsifiability, states that 'It is easy to obtain confirmations, or verifications, for nearly every theory – if we look for confirmations' (2000 [1963]:35).

One other possibility is through distributional factors; it can be hypothesized as a general principle that analogies are used as introductory mechanisms, given the results above, as a strategy for explaining abstract concepts in texts before using them to illuminate the consequences and import of this concept. Were this to be so, then when looking for distributed but non-global analogies (to establish that *WP* is indeed locally analogous), we may wish to examine the beginning of each chapter for analogous material. Doing this, it is immediately apparent that each chapter begins with a different analogical exchange, sometimes solely verbal (eg Chapter 1); in fact the Introduction states:

To make the transition for the new ideas in each chapter, I've begun the chapters with a brief story that isolates a key concept using more familiar metaphors and settings. I'm having fun with these stories, so go back to catch the references after you've read the chapter if you like.

(*WP*; Randall 2005:x)

These two findings make it possible to state that analogical domains are in use in the *WP* text, but they are not global. In common with *PW* above, this is most likely due to the broader nature of the text. Usefully, this allows us to place the texts on a cline, which is further elaborated in §3.iv below.

3.ii.4.c: *WP*: Summary

The analysis in this thesis aims to integrate computational and quantitative methods of obtaining semantic information with an existing tradition of qualitative analysis. It is therefore appropriate that these short quasi-quantitative analyses end with a text resistant to the wide-domain methodology outlined in this chapter. *WP* has analogical content, but it is not amenable to wide-domain analysis due to its local nature. This makes the corpus usefully broad, and leads to the analyses in Part II later in this thesis.

3.III: COMPARISON TEXTS

Before a final overview of the texts above, the three comparison texts are now analysed briefly, by contrast with the core texts above, in order to investigate the robustness of the analytical model proposed in this thesis. A reminder which may be necessary here: the three comparisons are an undergraduate-level number theory textbook (*C-NT*), a stylistics Masters' thesis (*C-AC*) and a popular science book (*C-GEB*) which straddles philosophy and computing and therefore contrasts in domain but not target readership with the popular physics/mathematics core texts above.

3.III.I: *C-NT*

The number theory textbook demonstrates the use of the above criteria for establishing analogical domains, as none of the domains returned by analysis were analogical in this text; all were relevant or noise. The analysis here required suspending the rule above, intended for popular science, that noise should only be determined by non-specialist knowledge (cf §3.ii.I.c.I.a and following above on *strings* and string theory). Instead, a knowledge of number theory terms (*field*, *matrix*, *ring*, *set*, etc) is necessary to determine noise. This is due to the differing discourse type of the university-level textbook as opposed to a text aimed at a general audience. Unfortunately, *set* (known to be the most polysemous word in *OED2*, although this is currently not the case in *OED3*; see Simpson 2007) is a key term in number theory, and produces significant amounts of noise. This is a useful check on the analysis above; should the system be useful in identifying key analogical domains then it should return no results for texts which are deliberately and systematically non-analogical.

A final point is that both noise and relevant domains here have a higher proportion of discourse connectives (*hence*, *since*, etc) than do other texts. These words are not included in the stopword corpus (due to their adverbial use in arguably non-function contexts; *four days hence*, etc) but their inclusion in the

key domains of this text could be used as a marker of a genre reliant on logical connections; this could aid future work using the *HT* to examine the use of adverbs in genre differentiation (see also, *inter alia*, Reppen *et al* 2002 and Biber 1995).

3.III.2: C-AC

This text was used simply as a check of the procedure during its development. A short list of key words and domains was created based on the text and my own knowledge of its contents, and this was compared against the returned key domains. This included *linguistics*, *psychology*, *rhetoric*, and *narrative* with regards to the thesis' topic, and *death*, *murder*, *lies*, *manipulation*, and *plot* with regards to the texts being analysed. The top 20 domains of the text which were statistically established as key are as follows:

HT Category	LL Score
Rhetoric	2073.5
Lack of confidence, distrust, suspicion	1476.6
Reader	1183.3
Interpretation	1137.0
Reading	1013.4
Publisher	1007.2
Branches/schools of linguistics	769.7
Materials of topic	727.5
Attestation, witness, evidence	673.6
Murder/assassination	664.2
Narrative/story	652.0
Narration	627.5
Memory	579.5
Evidence	466.0
Written text	426.4
Care/carefulness/attention	388.8
Newness/novelty	375.8
Speech/act of speaking	372.1
Fiction	367.4
Version of text	363.8

This provided a useful check of the category, as these domains overlapped substantially with the predicted list. Again, as expected, this text contained no wide-scale analogical domains.

3.III.3: *C-GEB*

C-GEB also had no key analogical domains *per se*. There were, however, two domains which were not immediately considered analogical, ostensibly relevant or noise; these were the domains corresponding to *crabs* and *turtles*. This requires analysis of the text and the concordance lines (not here reproduced) to discover that the analogical material in the text is represented by means of short narratives featuring characters taken from fables (Achilles and the tortoise, and a crab friend of the tortoise). These characters could broadly be described as analogical, although given the broad nature of popular science there is a strong argument to be made to create a narrative sub-category within the concept of analogical domains; many incorporate narrative elements as part of their strategy to mediate expository material.

The purpose of *C-GEB* was therefore to test the methodology on a popularized text that was not from the same area as those analysed in the core of the thesis. In terms of its subject matter, this text is different from the core texts as one which originates from the humanities areas of philosophy and musicology as opposed to the 'hard' sciences of mathematics or physics. It was hypothesized above in Chapter I that the need for analogy lessens the more concrete the discussion becomes; the results of this text support that, although much further research is needed to prove or disprove that general supposition. Nonetheless, the analysis supports the hypothesis that there are some systematic elements to be found in popular texts like this one, but an expansion of the methodology of this thesis into popular scholarly work beyond that of the hard sciences would require revisions to the procedure above, and perhaps the development of a schema of analogy.

3.iii.4: Overview

This comparison corpus was made up of three texts which were intended to contrast with the core texts with regards to their purpose; all of the texts, core and comparison, were in some way expository but differed from physical/mathematical popular science in one particular way (*C-GEB* was not physical/mathematical, *C-NT* was not popular science but still a scientific exposition, and *C-AC* was neither but still academic). The lack of analogical domains in these texts was expected and confirmed. *C-AC* provided a predictive check on the results of key domains, while *C-GEB* provided some further pointers with regards to the nature of analogy in different text genres. One further point here is that in contrast to the core texts, which all had core domains belonging to *OT The Physical World* (the core domain being the relevant domain with the highest LL score), *C-GEB* and *C-AC* differ.²⁵ *C-GEB*'s is *02.01.08 Understanding*, which mirrors its ostensive topic accurately. In a broad-stroke manner, the core domain (which was used to establish domain relevance, not for any other reason) appears to mirror the text's main topic. Importantly, none of the core domains were analogical, echoing the assertion above that analogy is subservient to the main thrust of the text.

3.IV: CONCLUSION

The analyses here have not aimed to use quantitative measures in such a way that they could replace qualitative analyses, but rather to supplement them. Two texts – *MP* and *FC* – were particularly amenable to wide-domain systematic analyses, whereas *PW* and (in particular) *WP* were not. Therefore they can be placed on a cline, from *MP* (the most directly globally-analogical text) through *FC* and *PW* to *WP*. All of them, with the exception of *WP*, have overarching key analogical domains, the use and distribution of which were

²⁵ For obvious reasons, *C-NT*'s core domain was *01.06.04.03 Mathematical number*.

demonstrated through the use of categorical distribution plots and lexical concordances. The existence of these analogical domains was established by a statistical procedure using fine-grained semantic material from the modern component of the *Historical Thesaurus*. All this data – the domains activated, the representative lexemes of these domains, the analogical concepts each text uses, and the distribution of these concepts throughout the text – can be used to ground and enhance the cognitive and stylistic analyses of selected extracts from these texts in the second part of this thesis.

The chapter began with the following research questions to address:

1. Does the technique described in Chapter 2 identify key domains in the corpus of popular science texts?
2. Does the technique identify key analogical domains in such texts?
3. What can be identified about the stylistic use of such key analogical domains across the whole text, and what, if anything, characterizes this use?

The analyses here have shown that 1 and 2 are indeed correct, for those texts which have key domains. The lists above go into more detail on question 2, and will be used in Chapter 4 below. Question 3 is a more interesting point. The analyses here have illustrated through the concordance analyses some characteristics of analogical domain use, which are here outlined:

Firstly, there exist lexical markers of analogical usage, which have been found to include *like*, *just as*, *as if*, *would be*, *picture* (as a verb), *imagine*, *analog/analogy*, *think of*, and further combinations of these (*just like*, *would very much be*, *imagine something just like*). These collocate with analogical referents in the concordance lines above. As a discourse function, they appear to mark areas where a reader is being explicitly told that the following text will be analogical and should not be interpreted wholly literally.

Secondly, analogical terms often collocate with others in the same analogical area following their introduction (so that *fabric* and *string* above were followed by *wire*, *thread*, *loop*, *patterns*, etc). Here, terms explicitly introduced by analogical markers then co-occur with terms from the same domain which are not surrounded by such markers. This suggests that both domains are being activated during the analysis, and that the concrete domain being used analogically is maintained in both its concrete and analogical forms, rather than being used to set up an analogy and then discarded once the mixed analogical domain is established. This effect, somewhat greater than ordinary topic coherence, provides evidence for the pluralistic and protean view of the semantics of discourse which is outlined in Chapters 1 and 2 above. As a linked point, there is often use above of the same term simultaneously in multiple domains (like the *wave* example in the *FC* analysis in §3.ii.1.a above), which maintains this dual-domain hypothesis.

Finally, when lexis from the concrete domain is used analogically, it is often accompanied by either an abstract premodifier (*cosmic fabric*, *mystical line*, *zeta landscape*) or an abstract postmodifying prepositional phrase (*of space and time*, *of the cosmos*, *of quantum reality*, *in the zeta landscape*) to explicitly relate the two domains wherever possible.

These characteristics are only those identified through concordance analysis of the key analogical domains selected above, and are not intended to be comprehensive. Nonetheless, the analysis here has provided an insight into the characteristics of the use of key analogical domains, by examining the context of use of the lexical items best representing those domains statistically identified as being key to the text in question. Together, they form an overview of the discourse neighbourhood of analogical domains in popular science texts, and are useful preparatory material for later analyses to come.

This concludes the empirical and statistical part of this thesis. The following chapter opens the second part, which begins to address the cognitive

consequences of the analyses which were begun here. It takes the empirical foundations that this part provided in order to undertake a cognitive analysis of significant extracts of the texts analysed above.

PART II: DISCOURSE-LEVEL

Language is a finite instrument crudely
applied to an infinity of ideas.

Stoppard 1972:63

CHAPTER 4:

A COGNITIVE APPROACH TO REIFICATION

4.1: INTRODUCTION

Following Part I, which empirically established the range of semantic fields in use during cross-domain mappings of popular science texts, the analysis undertaken in this part of the thesis addresses questions of local textual usage and of the semantic construction of analogical meaning at the level of relatively short text extracts, taken from the books analysed in Part I. The present chapter identifies appropriate extracts and then goes on to outline the cognitive linguistic analysis undertaken in Chapter 5.

The following analyses work within the semantic framework of Conceptual Integration Theory (also known as *conceptual blending*), supported by the cognitive approach to pragmatics known as Relevance Theory. By taking insights from both of these two approaches to meaning – one concerned with pluralistic semantic meaning construction, and one concerned with pragmatic principles employed by a reader/hearer to be best guided from protean meaning potentials to the writer/speaker’s meaning – this part of the thesis aims to demonstrate the complexity and variety of blends used within particular stretches of text, and to take steps towards integrating the empirical work in Part I with the cognitive linguistic analysis presented in this part to form a unified approach to multiple-domain mappings as a reification strategy with respect to discourse construal. Some developments are offered to, *inter alia*, Conceptual Integration Theory’s accounts of extensive multiplex blends, and a substantial blend creation procedure is proposed, which uses detailed empirical and linguistic criteria wherever possible to guide the description of conceptual integration, in contrast to the currently dominant intuition-based paradigm.

In terms of the overall argument of the thesis, this part extends and narrows the focus of Part I into a more explicitly cognitive analysis of construal, sacrificing the breadth of Part I for the necessary depth of analysis required to investigate a coherent textual construct.

4.II: 'KEY' EXTRACTS

4.II.I: OUTLINE

In the previous chapters, the four texts discussed in this thesis were analysed to discover their key analogical domains. Two of these texts had a series of domains which were used on a wide scale throughout the text, while the remaining two did not, instead being the subject of a hypothesis that they have domains of local analogical content, but not global as the former have. To establish appropriate samples for analysis in these chapters, the full corpus being much too large for the scope of this thesis, these two types of analogical content are handled differently below.

4.II.2: GLOBAL ANALOGICAL CONTENT: KEY DOMAIN EXTRACTS

The analysis in the following chapter requires a series of fairly short extracts. The length of these should ideally be no more than a thousand words, as the analysis below operates on the basis of lexical items and mental spaces, and is thus at a relatively fine-grained level when compared to the analysis in Part I. Four extracts totalling six thousand words in length would be a large enough sample size to yield an appropriate number of analogies while being much more manageable for the purposes of an analysis in a thesis of this size. Six thousand words would therefore represent one percent of the full 600,000 word corpus.

It is preferable for these samples to be chosen based on some relatively straightforward impartial criteria, as opposed to the common practice of an

analyst reading the texts and noting parts of particular interest. This is not intended to criticize this style of procedure, but is rather an extension of the assertion in Chapter 1 that a methodology which is grounded in digital and statistical measures is preferable, for the work in this thesis, to one which relies on analyst intuition. The lengthy process of Chapters 2 and 3 has been to identify texts and analogical material without a detailed knowledge of the text, in the tradition of digital humanities and corpus linguistics, and it is preferable to continue this as far as possible – until the cognitive linguistic analysis presented in the following chapters is undertaken, which necessarily requires an analytical process involving human intervention at a greater level than in Part I.

Therefore, the procedure outlined in this section aims to identify ‘chunks’ of text with high levels of analogical content for the three texts for which this is appropriate (*FC* and *MP*). As above, the lexical items most representative of the text’s respective analogical domains were taken as a proxy for analogical content itself, meaning the procedure discovers areas of the text where these items are most prevalent. A secondary consideration here is that for the textual ‘chunk’ to be as representatively analogical as possible, it should contain a mixture of these domains (and therefore representative lexemes).

Procedurally, each text was divided into smaller files, based on the numbers of lines in each original text file (representing a whole book). Each text has approximately fifteen to twenty words on a line, and lines were considered preferable to words as a metric for splitting the texts for a number of reasons: while it would have been possible to divide the text into smaller files depending on word counts, it was firstly simpler to do so with lines, secondly more appropriate as these files were intended to only be rough divisions, not needing word precision, thirdly preferable as paragraph breaks (and therefore topic shifts) coincide with a line ending, and finally more appropriate as a way of retaining line divisions to aid in later identifying subdivisions of textual units

distinguished by typography (such as extra blank lines between subsections, or ornamental flourishes acting as topic markers, for example asterisms (***) or horizontal lines). If each line contains approximately 15-20 words on average, then to get a very rough approximation of 500 words per divided file, a split of the text every 30 lines would be adequate.

This was carried out on all four texts using a simple Python program given in Appendix I as LineSplit.py, giving the following results:

Text	Total Word Count	Number of Divided Files	Words per Divided File
<i>FC</i> (Greene, <i>The Fabric of the Cosmos</i>)	192145	595	322
<i>MP</i> (du Sautoy, <i>The Music of the Primes</i>)	125980	328	384
<i>PW</i> (Kaku, <i>Parallel Worlds</i>)	130448	518	251
<i>WP</i> (Randall, <i>Warped Passages</i>)	151660	330	459

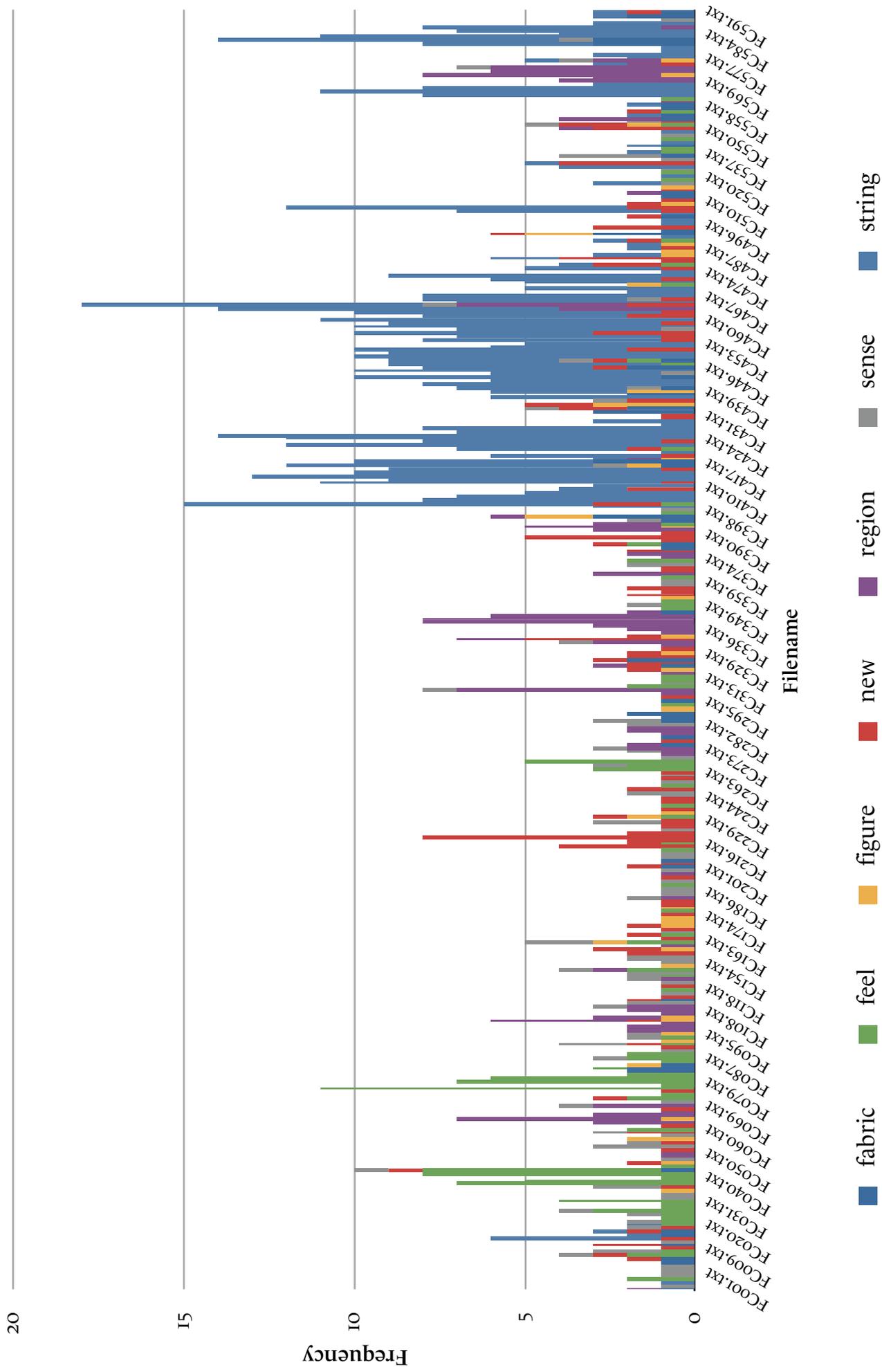
The resulting files were then loaded into Python using the NLTK PlainTextCorpusReader, a routine which takes a series of text files in a given directory as a single corpus. This allows searching of the corpus with results presented as a conditional frequency distribution, in this case summarized by filename. Searching for the lexical items instantiating the text's key analogical domains then gives counts of the occurrence of these lexemes in each 30-line divided file, and the results of these are given in Appendix III and summarized below. Each of the three texts in this section had clear clusters of analogical content, and so sections for analysis in later chapters were determined.

4.ii.2.a: Textual Clusters of Analogical Content: *FC*

The lexical items used here are the same as those outlined in §3.ii.1.c.1 above, namely: *string*, *fabric*, *feel*, *sense*, *new*, *region* and *figure*. Searching for these amongst *FC*'s divided files produces data which can be displayed in the following bar chart in Figure 5.

The horizontal axis represents the filename of the divided file (note that due to space restrictions only some filenames are shown, even though all the data

Figure 5: FC Analogical Textual Clusters



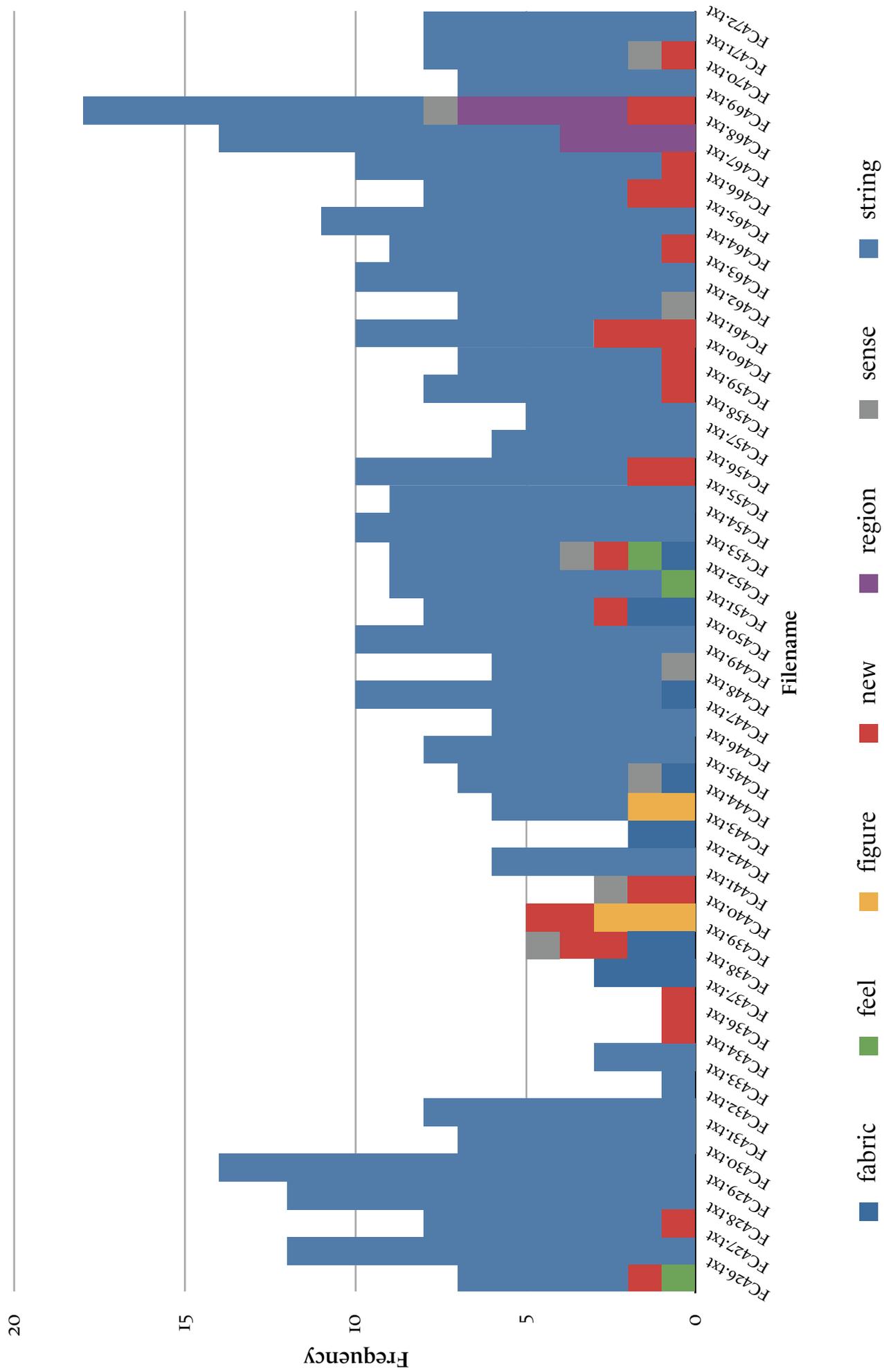
appears on the chart), and the vertical axis shows how many occurrences of each search term can be found in each file. It is immediately apparent that the *string* lexeme is here heavily represented in certain clusters, but as discussed in §3.ii.I.c.I, its use as a technical term makes its use here more complex than it would otherwise be. There are a number of peaks (for example, *feel* in the 80s filename range, and *new* in the 220s), but large clusters with mixed lexemic content occur later in the text, most notably in the 440s range. Until this point, there are no significant clusters (ie above two occurrences per file) of at least four of the seven analogical domains in the text. As stated above, this is a useful criterion as it means each textual ‘chunk’ for analysis is as representatively analogical as possible, rather than simply a heavy instantiation of one particular analogy.

This region around the 440s is represented in more detail in Figure 6. The cluster here between 440 and 445 represents a range of mixed analogical content and spans 5 files, giving 1610 words. This is an adequate range to then be analysed separately for topic divisions, using the criteria above.

Taking this extract from the text, and going back to the formatted version of the text rather than the plain text used for corpus analysis above, provides clear typographical cues as to textual divisions. Four-fifths of the way through this extract, there is a chapter boundary where chapter 12 ends and 13 begins, and similarly, near the end of the extract, there is a subsection boundary, typographically marked by a heading. These topic shifts can be used to delimit the text to a single 968-word chunk, which is acceptable as a text for analysis in the following chapter.

The following sections undertake the same attempt to find a coherent textual extract of analogical clusters, but the explanation is abbreviated compared to the more detailed outline given in this section.

Figure 6: FC Analogical Textual Clusters (extract, files 426-472)



4.ii.2.b: Textual Clusters of Analogical Content: *MP*

The key lexical items for *MP* were outlined in §3.ii.3.c.1 above, and are *way*, *far*, *line*, *level* and *point*. These occur in *MP*'s divided files with the distribution shown in Figure 7.

Ignoring the single-term spike of *line* in 224, there are two clusters here, centring around 97 and 277 respectively. From the graph, it is apparent that the cluster at 97 is more mixed than the one at 277, as the latter consists mainly of the lexeme *level*, whereas the 97 cluster has a wider range of analogical items, as can be seen from Figure 8, which centres on that region.

There is a chapter break just after the beginning of this region shown on Figure 8, partway through file 85. This chapter ends part-way through file 102, which gives a chapter of 6527 words – much too long for the purposes of this thesis. Taking instead the initial part of this section would be more manageable, and there is a paragraph break 1004 words into the chapter which functions well as a divider. (It is worthy of note here that the landscape analogy outlined in §3.ii.3.c.1 *et seq* above is heavily represented in this sample, which is to be expected given the method of extract selection.)

4.II.3: LOCAL ANALOGICAL CONTENT

Both *PW* and *WP* were discovered in Chapter 3 to have a lack of global key domains, which that chapter set out to investigate. This section therefore aims to find appropriate sets of examples of local analogical usage (*local* here is again used in the sense of a domain being analogically employed within a short stretch of text, and thus contrasted with the textually-global domains in §4.ii.2 above). By this it is meant that some extracts from these texts should be found which are relatively analogical, for analysis alongside those with key text-wide analogical domains.

To do this, the observation in §3.iv above that there are lexical markers of analogical content identified in the concordance lines throughout §3.ii gives a

Figure 7: MP Analogical Textual Clusters

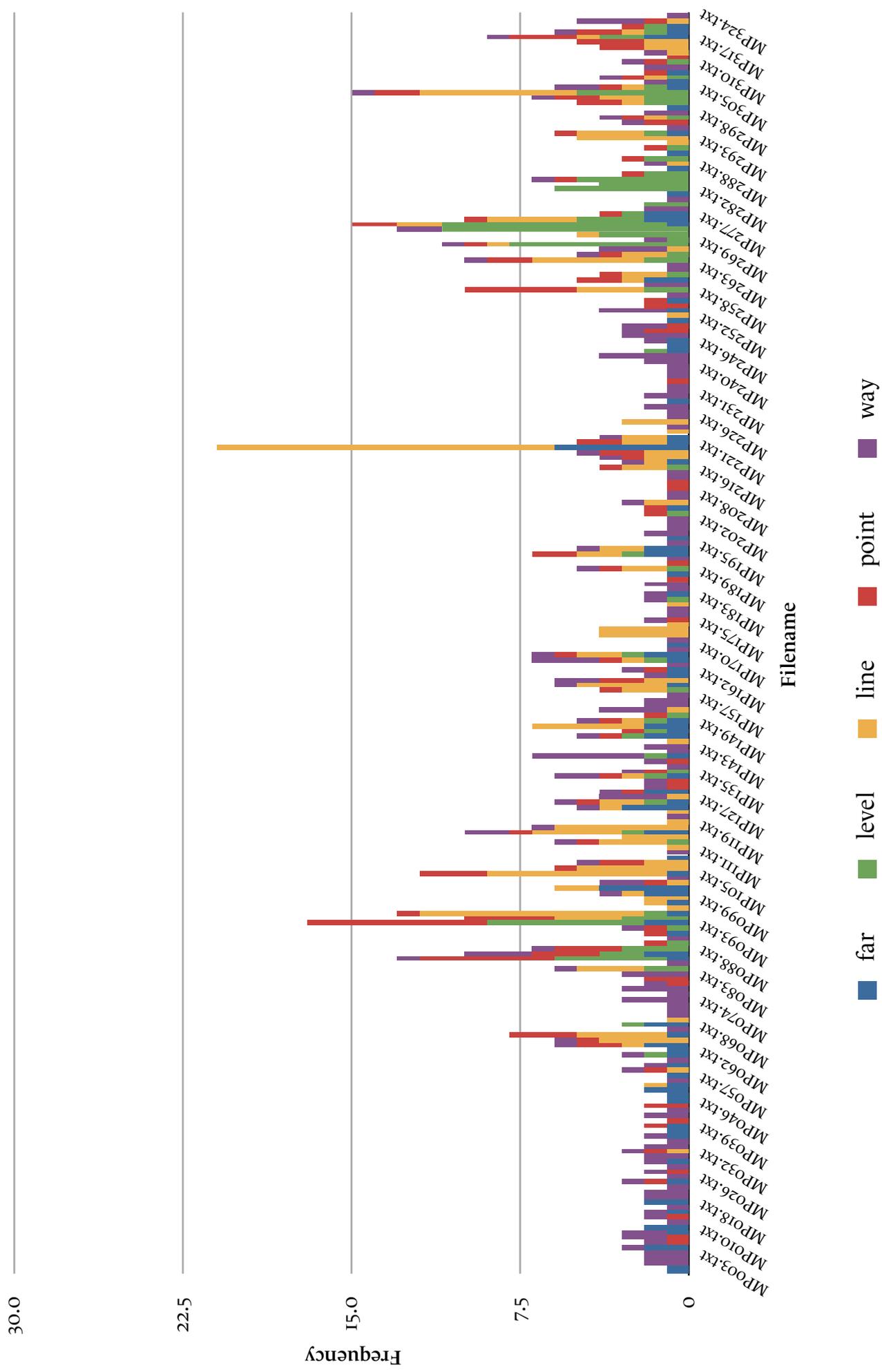
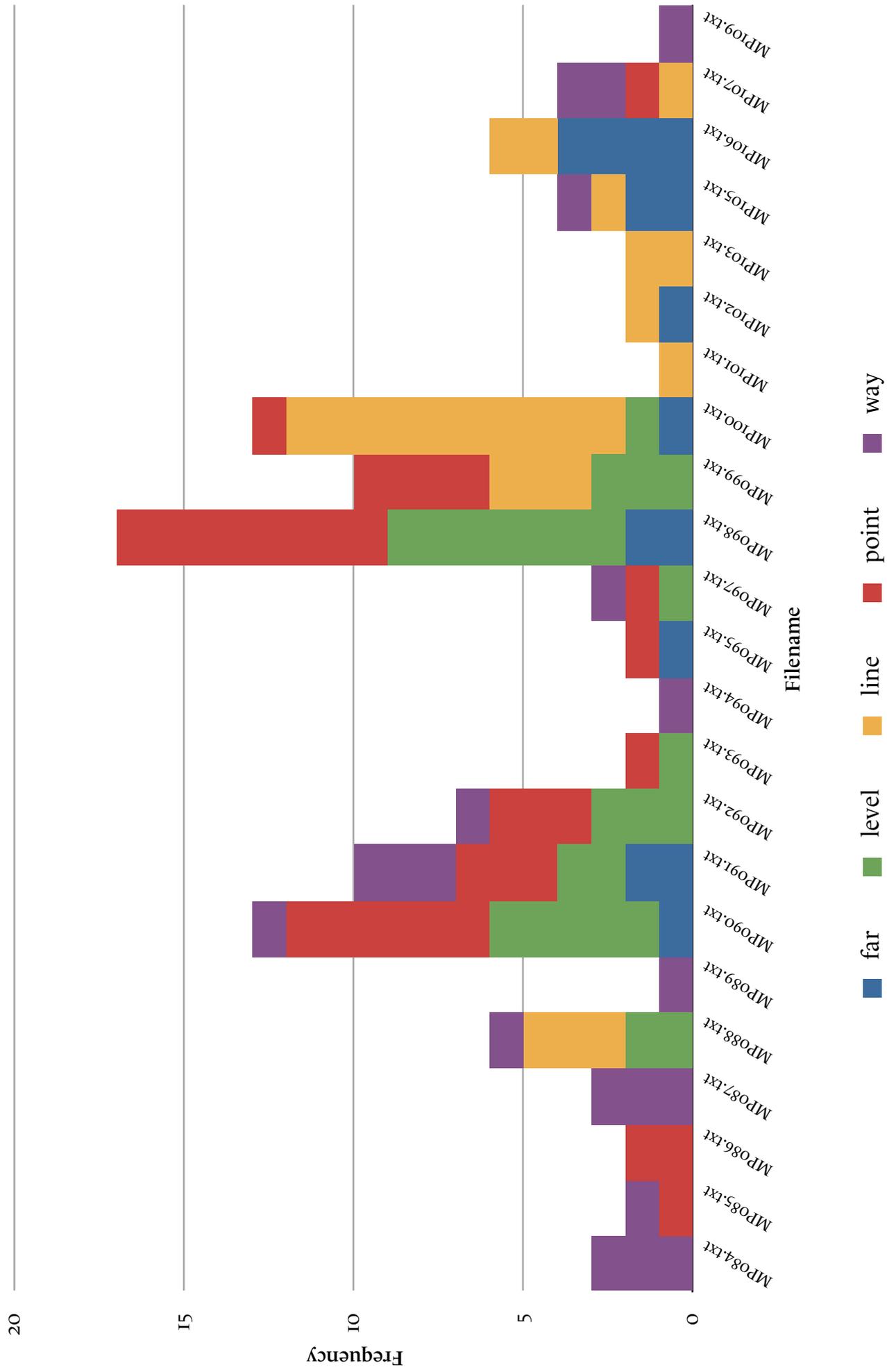


Figure 8: MP Analogical Textual Clusters (extract, files 84-109)

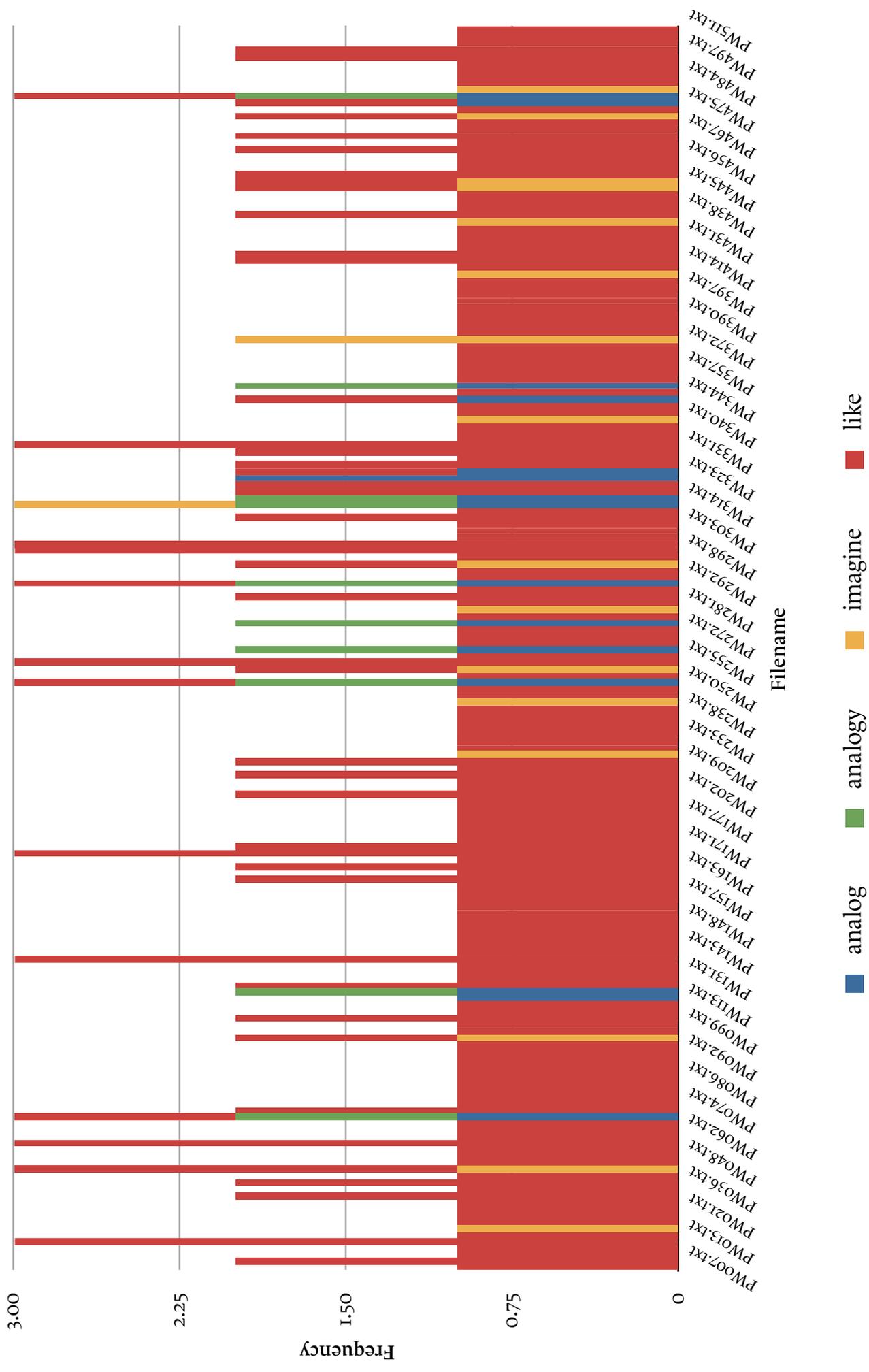


way of searching through the divided text files for likely analogical content. These markers were listed as *like*, *just as*, *as if*, *would be*, *imagine*, *analog*, *analogy*, and *think of*. Although not intended to be comprehensive, this list functions as a usable proxy for analogical markers in the texts under consideration, with the caveat that *analog* is particularly polysemous in the scientific field (as an antonym of *digital*), and *like* is a polysemous term in English. Only the single-word lexemes are used here, as a consequence of the lemmatization required for conditional searching.

4.ii.3.a: A Sample of Local Analogical Content: *PW*

Searching for these markers in the divided corpus of *PW* gives the graph shown in Figure 9. This graph looks very different to the others above, with a much shorter y axis representing the relative infrequency of these lexical markers compared to the lexemes above representing key domains. The largest mixed cluster here is in the region of 312 onwards, and there is a section break in 311 which functions as a useful starting point. From here, there is a further section break approximately one thousand words later which functions as a final delimiter. Reading this extract to confirm the hypothesis that this approach will find material of use for later analysis finds multiple markers of analogical material, including topic introductions like ‘By crude analogy, think of...’ and ‘is the universe a hologram? [...] is the universe a computer program?’. This is sufficiently analogical, and although no claims are made here that it is the densest cluster of analogy, it is taken to be appropriately representative of analogical content in the text and so gives a 1,017-word extract for later analysis.

Figure 9: PW Analogical Textual Clusters



4.ii.3.b: A Sample of Local Analogical Content: *WP*

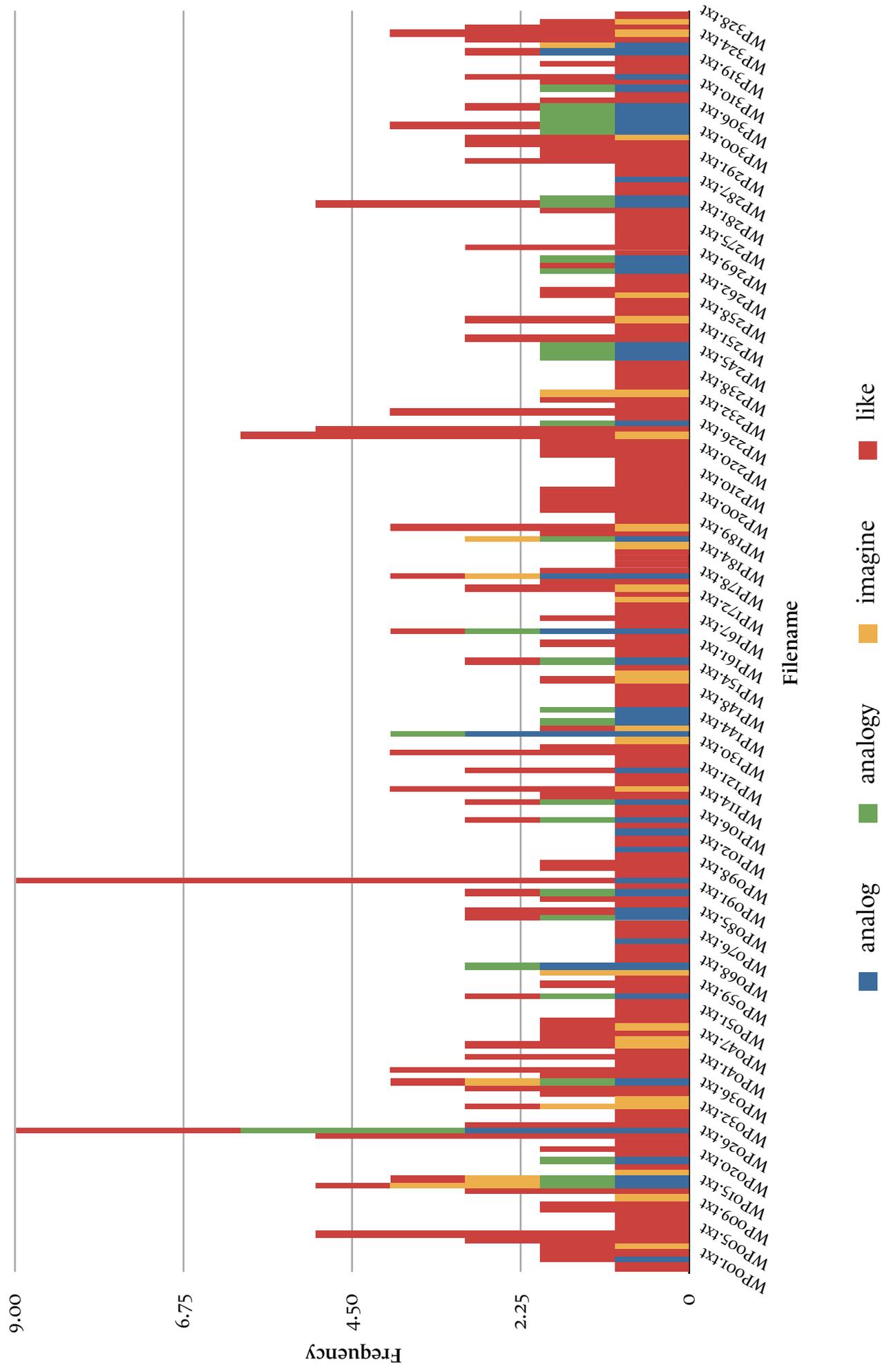
WP, as a text which was particularly resistant to the method above of searching for global analogical content, nonetheless has a reasonable number of analogical markers spread throughout the text – again, with the caveat that this short list of markers is not in any way comprehensive. A bar chart of these is shown in Figure 10.

The largest mixed cluster here is in the 15 to 26 range, which is the only cluster of any appropriate size to contain all four analogical markers. Beginning at the section break in 17, there is again another section break after a little under a thousand words, giving a 977 word sample for analysis.

4.II.4: SUMMARY

This section has identified four extracts from the texts analysed in Part I, firstly by looking for clusters of key analogical domains, for those texts which have them, and secondly by looking for clusters of analogical markers, for those texts with non-global analogical content. These texts are reproduced in Appendices II.2-5. The following sections now outline the mode of analysis of the remainder of the thesis.

Figure 10: WP Analogical Textual Clusters



4.III: QUANTITATIVE TO QUALITATIVE

The analyses here have not aimed to use quantitative measures in such a way as to replace or challenge those qualitative analyses normally undertaken in cognitive stylistics, but they have rather been intended to supplant the often impressionistic procedure of choosing material for analysis. The system used above to choose such material, mainly through the combination of existing corpus techniques, computational linguistic methods, and general data mining principles, aimed firstly to independently²⁶ prove the existence of systematic analogical material in the texts selected for analysis and secondly to provide evidence of clusters of such material with sufficient density that (by comparison with the remainder of the text), they could be confidently used to represent some of the densest areas of analogical material in the whole text.

It is at this point in this thesis that the quantitative ‘distant reading’ approach above dovetails with the following qualitative cognitive stylistic approach (a return, to expand Franco Moretti’s 2000 analogy, to the ‘close reading’ of textual interpretation from which stylistics as a discipline emerged). At this stage, quantitative methods have provided, from a corpus of over six hundred thousand words, extracts of approximately one thousand words for analysis. Wherever possible, the qualitative analysis below will use support from the quantitative methods employed above, most particularly in the use of lexicological data of the *Historical Thesaurus of English*, but the style of analysis here changes substantially.

²⁶ ‘Independently’ here is a problematic term; it is intended to mean independent of any one individual’s impressions of a text gained through the reading process, rather than wholly divorced from any possible interference by an analyst. Indeed, the early choice of cognitive semantic analyses as the eventual end output of the entire process of this thesis has influenced the word-based analytic procedure chosen and the size of the eventual extracts for analysis (focusing on 1-gram/unigram units, in corpus linguistic terminology). Beyond this, it is axiomatically accepted in the present thesis that no analytic procedure in linguistics can be wholly independent of an analyst’s influence, but merely that it can strive to give an evidence base which is not dependent wholly on that analyst’s impressions.

In particular, the analyses below are within the framework of Conceptual Integration Theory. This is a relatively new theory within cognitive linguistics and stylistics, and one with multiple possible analytical procedures, each devised by scholars with differing methodical requirements. The present thesis is no exception to this diversification, and indeed contributes to it by introducing a method of carrying out conceptual blending analyses supported by the data of the *Historical Thesaurus*. Beyond this, however, it does not propose any radical changes to the core methodology as set out by its primary theorists.

The remainder of this chapter therefore outlines the major precepts of conceptual blending, and introduces in outline form the modifications made for the analyses carried out in this thesis. It is a relatively long exposition, but is considered necessary because of the shifting nature of the theory's present application. The outline below describes the position taken on Conceptual Integration Theory within this thesis, and in so doing states the aspects of the theory which it follows.

4.IV: CONCEPTUAL INTEGRATION THEORY

Conceptual Integration Theory²⁷ (hereafter CIT) is 'a theoretical framework for exploring human information integration' (Coulson and Oakley 2000:176), or 'a set of noncompositional processes in which the imaginative capacities of meaning construction are invoked to produce emergent structure' (Coulson 2001:115). In more detail, Gilles Fauconnier and Mark Turner, its leading theorists state:

Conceptual integration - 'blending' - is a general cognitive operation on a par with analogy, recursion, mental modeling, conceptual categorization, and

²⁷ The theory has also variously been known as *conceptual blending* (which is treated as synonymous with *conceptual integration* in the literature), *mental space blending*, and, more rarely, *the network model of conceptual integration*.

framing. It serves a variety of cognitive purposes. It is dynamic, supple, and active in the moment of thinking. It yields products that frequently become entrenched in conceptual structure and grammar, and it often performs new work on its previously entrenched products as inputs. Blending is easy to detect in spectacular cases but it is for the most part a routine, workaday process that escapes detection except on technical analysis. It is not reserved for special purposes, and is not costly²⁸.

In blending, structure from input mental spaces is projected to a separate, 'blended' mental space. The projection is selective. Through completion and elaboration, the blend develops structure not provided by the inputs. Inferences, arguments, and ideas developed in the blend can have effect in cognition, leading us to modify the initial inputs and to change our view of the corresponding situations.

(Fauconnier and Turner 1998:133)

These claims may appear somewhat strong, however the theory has developed from its origins in the 1990s into an extensive cognitive approach to the construction of meaning in natural language from a domain-based perspective. While used and tested in some psychological work, in recent years the theory has been used mainly as a linguistic approach to dynamic meaning construction. It contains within its framework an account of metaphor theory in the Lakovian mode²⁹, and is (with most of cognitive linguistics) supple enough theoretically to incorporate insights made in other areas of the field.³⁰ With regards to this thesis, it is asserted that a sufficiently complex, theoretically broad and domain-based approach to natural language semantic processing is ideal for answering the present key research question of how to describe the cognitive strategy of reifying discourse construal in popular science. Conceptual integration fits this requirement better than any

²⁸ 'Costly' here appears to be used in the sense of requiring extra mental processing time. [my note].

²⁹ CIT actually 'presupposes' cognitive metaphor theory, in the words of Grady *et al* 1999.

³⁰ See, as also mentioned above, the 'blending typology' of Line Brandt (2002).

alternative cognitivist theory; conceptual metaphor theory in either the Lakovian mode (eg Lakoff 1987 and Lakoff and Johnson 1980 and 1999) or the European mode (eg Steen 2007, Semino 2008 or Low *et al.* 2010) is not as concerned as conceptual integration with the nature of domain mappings, and also has a frequent preoccupation with existing focused conceptual patterns, rather than emergent properties. Other theories, such as Vyvyan Evans' LCCM (Lexical Concepts and Cognitive Models) theory (2009), are too new to have had their ontological status and analytic appropriateness adequately tested, while those cognitive stylistic theories focused on literary narratives are not appropriate to adapt into exposition-focused models. Some further points on this topic are made in passing below.

4.IV.I: LITERATURE

The most notable work on CIT is Fauconnier and Turner's 2002 *The Way We Think*, which summarizes integration theory as it had developed to that point and extends it, incorporating some major theoretical changes from previous works (such as Fauconnier and Turner 1998). One particular feature of Fauconnier and Turner's work is that Turner in particular popularizes the theory outwith academic circles; his book addressed to social scientists (Turner 2001), for example, explains blending in terms of 'notions' and 'influences' rather than using the technical terms he introduces in other articles. So while *The Way We Think* is thus far the 'core' book on the theory, it has a strong focus on accessibility, occasionally to the detriment of detailed argument (see, for example, Brandt 2005 for a specific challenge of an important technical argument skimmed over in Fauconnier and Turner 2002). By contrast, Coulson 2001 is more explicitly psycholinguistic, as are the articles by Grady *et al* 1999, Coulson and Oakley 2000, Grady 2005 and Fauconnier and Turner 1998. Much of the following discussion is based on those sources in addition to (and occasionally in lieu of) *The Way We Think*, although quotes are taken from

that book where possible, as the definitive statement thus far of the theory by its originators.

4.IV.2: THE PROCESS OF INTEGRATION

The dynamics, sources and outcomes of this process of integration are outlined below. An important starting point is that the description of blending by an analyst is assumed to be a representation of the mental processes undertaken in the construction of the blended concept;³¹ so any account of conceptual integration will take the form of a procedural exposition of the mental process involved in the reader/speaker's creating of the end product – known as the *blend*.

An instance of blending involves at least two mental *input spaces*, much like the domains used in conceptual metaphor theory. Unlike conceptual metaphors, however, CIT does not use 'source' or 'target' domains, instead having two or more separate mental spaces (see §4.iv.3 below on mental spaces), constructed *ad hoc* and which contain *elements*, some of which share similar features (which are realized in the text). The input spaces are then compared and *connections* are made between counterparts in each space. This gives rise to a third space, the *generic space*, which represents what the inputs 'have in common'. The final space is then constructed by projecting elements from the input spaces into a *blended space* (usually termed the *blend*, as above), where the elements which have been mapped through connections as having something in common (and are represented in the generic space) are fused into a single element in the

³¹ 'In the case of blending, the effects of the unconscious imaginative work are apprehended in consciousness, but not the operations that produce it [...] The dynamic web of links between blend and inputs remains unconscious. What is registered consciously is the encounter in the blend and the "consequent" alignment between the two concepts', and, somewhat more obliquely, '...it is in the nature of consciousness that it gives us effects we can act on, and these effects are correlated with the unconscious processes. In the case of meaning, the apprehension of these effects typically induces us to reify meaning. Consciousness sees an effect and reifies the effect to produce a cause: I see a cup, and, in the folk theory, the reason I do is that there is a cup that causes me to see a cup. In the same way, I hear a sentence and I "see" a meaning for it...' Fauconnier and Turner 2002:56-7.

blend, while some other elements which cannot be fused are instead imported into the blend as-is. For these fused elements, the ‘prominent counterparts from the input spaces project to a single element in the blended space [so that] a single element in the blend corresponds to an element in each of the input spaces’ (Grady *et al* 1999:114). The procedure can be recursive if required, and importantly for this thesis, any complex process of integration can result in multiple interweaving blends (see below).

4.iv.2.a: The Nature of the Blend

The major claim of CIT is that the integration of these spaces not only draws upon the inputs but actively creates the blend, which is an emergent structure whose setup is different from each of the inputs and which could not be predicted by either in isolation. Fauconnier and Turner describe three processes which contribute to this emergent meaning, with reference to an old puzzle.³² In short, the puzzle describes a situation where there are two input spaces of the same monk walking a single route twice, once upwards (T_1) and once downwards (T_2). The blend fuses the route but maintains the monk at T_1 and the monk at T_2 as separate entities in the blend (as not to do so would eliminate salient information about their direction of motion, amongst other things, and so they cannot be fused). The emergent structure is thus very different from the inputs, as Fauconnier and Turner explain:

First, *composition* of elements from the inputs makes relations available in the blend that do not exist in the separate inputs. In the blend but in neither of the inputs, there are two moving individuals instead of one. They are moving in opposite directions, starting from opposite ends of the path, and their positions can be compared at any time of the trip, since they are traveling on the same day [...]. Second, *completion* brings additional structure to the blend. This structure of

³² A Buddhist Monk begins at dawn one day walking up a mountain, reaches the top at sunset, meditates at the top for several days until one dawn when he begins to walk back to the foot of the mountain, which he reaches at sunset. Make no assumptions about his starting or stopping or about his pace during the trip. Riddle: Is there a place on the path that the monk occupies at the same hour of the day on these two separate journeys?

two people moving on the path can itself be viewed as a salient part of a familiar background frame: two people starting a journey at the same time from opposite ends of a path. Third, by means of *completion* [*sic*], this familiar structure is recruited into the blended space. At this point, the blend is integrated: It is an instance of a particular familiar frame, the frame of two people walking on a path in opposite directions. By virtue of that frame, we can now run the scenario dynamically: In the blend, the two people move along the path. This ‘running of the blend’ is called *elaboration*. Running of the blend modifies it imaginatively, delivering the actual encounter of the two people. This is new structure: There is no encounter in either of the input mental spaces, even if we run them dynamically. But those two people in the blend are projected back to the ‘same’ monk in the two input mental spaces. The meeting location projects back to the ‘same’ location on the path in each of the inputs [...]

(2002:42-44, their italics)

These three features create the blend as a comprehensible and distinct space; to summarize, *composition* composes ‘elements from the input spaces to provide relations that do not exist in the separate inputs’ (such as two men on the same mountain), *completion* is the use of schematic knowledge so that a ‘minimal composition in the blend is often automatically interpreted as being a richer pattern’, and *elaboration* is metaphorically explained as if the blend was a simulation, where the mind ‘runs’ it ‘imaginatively according to the principles that have been established for the blend’ (quotes from Fauconnier and Turner 2002:48).

Conceptual integration’s central claim, as it concerns language and its interpretation, is therefore that there are cognitive explanations of linguistic phenomena made possible by describing those phenomena in terms of the blends they result in.

4.IV.3: MENTAL SPACES

Important here within the CIT approach is that mental spaces are temporary representational structures constructed at the point of utterances and responsive to contextual factors (Fauconnier and Turner 1998, 2002 *passim*). But conceptual blending arose at least in part from Fauconnier's earlier work on what he also termed *mental spaces*, theoretical constructions of entities, relationships and processes which are intended to describe referential structure (Fauconnier 1985). In particular, the theory aims to describe reference and co-reference in four 'types' of situation; domain spaces, time spaces, space spaces and hypothetical spaces (which refer respectively to one field of activity, a temporal location, a geographical location or a conditional, hypothetical or unrealized situation). One of mental space theory's advantages was its ability to partition information within a discourse – to create, say, both a hypothetical space and a real-time space in a sentence such as *I could have an English accent if I had been brought up in Essex, but I'm really from Glasgow*. Such partitions are then linked by *connectors* between the *counterparts* (those elements which have a correspondence between the partitions), such as the 'I' in the example above, which exists both in the real-time space (which is known as the *base space*) and the hypothetical space. Within the hypothetical space, the 'I' has a relation of having an English accent; within the real-time space, the 'I' has a relation of being from Glasgow (and, by implicature, having a Scottish accent). These partitions are created through either implicit knowledge that an alternative space is being discussed (a deliberately rather vague requirement in mental space theory) or through explicit *space-builders* (adverb phrases, prepositional phrases, conjunctions or what Fauconnier calls subject-verb 'complexes' – which appear to generally be what Hallidayan linguistics would term mental or verbal processes such as *X thinks, X believes, X argues* where *X* is a noun phrase).

Although Fauconnier's focus is on language-based examples for the majority of his uses of mental space theory, it is explicitly not a solely linguistic approach; the claim, as with most of cognitive linguistics, is that this is a linguistic form of

a general cognitive process (see also Coulson 2001:24 for references to work in behavioural psychology). Conceptual integration, although growing out of mental space theory, makes surprisingly little reference to it. For example, Fauconnier and Turner, in *The Way We Think*, simply define mental spaces as ‘small conceptual packets constructed as we think and talk, for purposes of local understanding and action’ (2002:40). They make some neural claims (‘mental spaces are sets of activated neuronal assemblies’, *ibid*) but otherwise appear not to promote the prior approach (spaces within CIT are described in a four-page summary in *The Way We Think*, which makes no reference to Fauconnier’s earlier work – 2002:102-106). The four ‘types’ of mental space above are still imported into CIT, but without the same explicit terms (2002:40ff). The influence of mental spaces on CIT is unmistakable, despite Fauconnier and Turner’s lack of explicit reference; the more overtly linguistic works on CIT (such as those mentioned above) generally include some reference to mental spaces, either in the context of explicating CIT or to clarify parts of the blending ‘approach’ with regards to the origin of certain aspects of Fauconnier and Turner’s theory. Fauconnier and Turner state that ‘input spaces and blends under construction recruit structure from more stable, elaborate, and conventional conceptual structures’ (1996:115). These stable structures from which spaces ‘recruit’ information can be assumed to be *schemata* (or *frames*, for some authors), cognitive structures representing generic knowledge within the mind (which is also implied by Coulson 2001:271-2). This lack of explicit reference between the two means that for the purposes of the analysis in this thesis, the term *mental space* will be interpreted as a domain or dynamically-instantiated schema (see Emmott and Alexander 2009/2010), without detailed reference to the earlier theory.

4.IV.4: APPLICATION

The use of CIT in this thesis is justified by the way in which it handles multiple concurrent input domains. However, as with many cognitive linguistic

theories, the detailed application of the theory to a particular text or expression can prove difficult should there be a requirement for an analysis to be detailed and (ideally) methodologically reproducible, as in this thesis. The following section is an attempt to outline the application and use of conceptual integration in some detail; readers experienced with the theory may wish to skim the following sections, although attention is drawn to the relevance of popular science reification to the sections titled *Compression* and *Overarching Goals* below.

4.IV.5: THE CREATION OF A BLEND

4.iv.5.a: Selection

It is not immediately apparent from the literature how the construction of input spaces and the movement of information from these to create a blended space functions. The retrospective quality of CIT (see §4.iv.8, below) makes it appear that the analyst has to somehow pre-construct the blended space from the text under consideration, and then retroactively identify which features of the input spaces are taken into the blend. While the theory states that input spaces are used by the mind to create the blend, the analyst must instead create the input spaces themselves *from* the blend. How does the mind then non-retroactively construct blended spaces?

Fauconnier and Turner are clear on this point, if characteristically a little light on references to other research. They state that as blend-construction is an ongoing, emergent, network-based process, it is also, by necessity, one which involves revision and changes to the construction during the process. Their explanation is worth quoting in full:

When we see the final integration network with all of its connections in place, it may look like a *tour de force* showing the mastery of its creator in selecting just the right projections [...] There is always extensive unconscious work in meaning construction, and blending is no different. We may make many parallel attempts to find suitable projections, with only the accepted ones appearing in the final

network. As we project to a blend, we are also working on the entire network, and we may, for instance, recruit new structure to the inputs precisely to make it available for possible projection to the blend. After the fact, it looks as if that structure was in the inputs to begin with, and as if building a network is a sequence of discrete operations; but that appearance is misleading. Input formation, projection, completion and elaboration all go on at the same time, and a lot of conceptual scaffolding goes up that we never see in the final result. Brains always do a lot of work that gets thrown away.'

(2002:71-72, their italics)

(We may also note at this point that this heavily modalized and rhetorical passage highlights the fact, implicit in most cognitive linguistics, that CIT is a *model* of meaning construction, which only makes an effort to account for the end result and not the detailed mental and/or neurological processes which occur. This is not unique to CIT, but is worth emphasising here for the sake of clarity.) The later point is made that some blends are generic within a culture, so that a generic blend is 'a general template widely applicable to specific cases' (2002:72). These *blend templates* (Fauconnier and Turner's term) therefore make blend construction easier if one or more input spaces are socially defined – 'the general form of the projections and the completion are specified in advance and do not have to be invented anew' (*ibid*).

4.iv.5.b: Compression

Ungerer and Schmid (2006:260) term compression 'the ultimate goal of the whole blending process', while Fauconnier and Turner state 'One of the most important aspects of [human beings'] efficiency, insights and creativity is the *compression* achieved through blending' (2002:92, their italics). Ungerer and Schmid continue:

The crucial effect of compression is that the conceptual complexity of the inputs from several sources is reduced considerably. A newly integrated and unified

conceptual structure emerges that is cognitively manageable and thus has [...] 'human scale'.

(2006:260)

Although their definition of compression is somewhat dependent on ostensive examples, Fauconnier and Turner use the term to refer to the condensing or reduction of relationships which exist between the input spaces. The canonical example above, the monk travelling up the mountain and back down, shows, in the blend, a compression of time – the blend places the monk's existence at two different times into a single mental space, eliminating the temporal distance between the monk at T_1 and the monk at T_2 . This particular blend therefore compresses one axis – of time – to represent the men being separated only in space; 'running' the elaborated simulation (to use Fauconnier and Turner's term – an alternative might simply be *imagining the situation*) thus requires the reader/hearer to consider the action of two men walking towards each other, rather than two discrete spaces separated by time and also going forward in time internally. This is what Fauconnier and Turner term **human scale** (202:312), which is core to the theory, and can be characterized as a situation which is easily processed, being based on simple, familiar (and perhaps even embodied) situations. Fauconnier and Turner again describe this important concept only by example (*ibid*):

The most obvious human-scale situations have direct perception and action in familiar frames that are easily apprehended by human beings: An object falls, someone lifts an object, two people converse, one person goes somewhere. They typically have very few participants, direct intentionality, and immediate bodily effect and are immediately apprehended as coherent.

They further state that the human scale 'is the level at which is it natural for us to have the impression that we have direct, reliable, and comprehensive understanding' (2002:323). A recent draft of an article by Turner (2007), available online, contains the rather more detailed but still not ideal definition:

Cognitively modern human minds are built to conceive of scenes that are at human scale. At human scale, we operate within certain ranges of space and time. We partition our sensory fields into objects and events, where some of those objects are agents. We engage with a few agents in patterned activity. We interact with objects. We detect, acquire, and manipulate objects, often as instruments for action. We eat, we move, we fight, we mate, we procreate. That is pretty much what we are built for. In one sense, it is what we are.

The obvious consequence of this for the present thesis, although not explicitly stated in the literature thus far, is that the reduction of complex situations to ones which are at human scale, and thus more easily apprehended, is a clear example of the sort of explication and reification which characterizes popular science. As outlined in the Introduction (Chapter I above), if we take the purpose of popular science to be an ‘interpretation’ of specialist literature for a general audience, then the use of human scale in the field would be a particular strategy to give a reader ‘global insight, human-scale understanding, and new meaning’ (Fauconnier and Turner 2002:92) of science. For now, this thesis will use ‘human scale’ as an equivalent for the concretized outcome of the process of reification, and ‘achieving human scale’ and ‘reify’ as synonyms. (See also *Overarching Goals* below.)

4.iv.5.c: Mappings

Continuing with the explication of CIT as it is used here, compression (the process of gaining human scale) operates on relationships between input spaces. Fauconnier and Turner define certain of these relationships in detail, terming them **vital relations** due to their ubiquity (2002:92). Relations in use are divided into two categories: **outer-space** (between two spaces, eg between two input spaces) and **inner-space** (within a single space, eg within a blend), although they are not distinguished in the general case, only when used in an analysis. For later reference, these relationships are summarized in the table in Appendix II.I, with quotes from Fauconnier and Turner 2002 next to the page references of the explication of each. Characteristically, Fauconnier and Turner

once more give only ostensive, example-based definitions (and these heavily imply that the relations are to be interpreted as common-sense terms). The names of vital relations are capitalized when used in a technical sense, although Fauconnier and Turner are not consistent about capitalising the term *vital relations*.

While the concepts here described as vital relations appear, on first glance, to be ‘natural’ terms for links between similar spaces, the problem for any analyst encountering *The Way We Think* will be immediately clear; these connections often have to rely on intuitive labelling and by making a link with the ostensive definitions laid out here and uses of each term in the literature.

4.IV.6: GOVERNING PRINCIPLES

To conclude this brief overview, Fauconnier and Turner introduce, in the final chapters of *The Way We Think*, an important set of governing principles for blending.³³ These take the form of *Overarching Goals*, *Governing Principles for Compression*, *Other Governing Principles* and *Constitutive Principles*. The *Constitutive Principles* are those which establish blends, as outlined above (connections, the generic space, blending, selective projection, emergent meaning, composition, completion and elaboration). For reasons of space, the remainder are only briefly summarized below.

4.iv.6.a: Overarching Goals

Fauconnier and Turner have one ‘overarching goal’ for all the principles outlined above: to achieve human scale. This leads to five subgoals, as achievement of human scale ‘often requires imaginative transformations of elements and structure in an integration network as they are projected to the blend’ (2002:312). These are listed by Fauconnier and Turner (*ibid*) as: to compress what is diffuse (either shortening the scale of a single relation, or compressing one or more relations into one another), to obtain global insight,

³³ These are often identified as explicit counters to the claim of the theory’s overapplicability (see §4.iv.8 on criticisms below).

to strengthen vital relations, to ‘come up with a story’ (on the assumption that the human scale is typically characterized by narrative events), and to go from Many to One (to fuse as many significant agents/events as possible; the reason for the capitalization is unclear).

These principles not only form CIT’s overarching goals, but can be seen as initial guiding factors for the description of the discourse strategy of the reification of abstractions in the texts this thesis is concerned with. This point is returned to in the following chapter, but for now the most significant point is that the goal of ‘obtain[ing] global insight’ is characterized as giving ‘mastery of a diffuse network, which creates a feeling of global conceptual mastery and insight’ (2002:324), and this mirrors the ostensive explicatory goals of the popular science genre (see further Chapter I above). Also, Fauconnier and Turner note with some significance that these subgoals are not independent of one another; ‘Compression is a way to achieve human scale, and by the same token achieving human scale will produce compression. Strengthening vital relations is also part of achieving human scale, and a scenario at that scale typically involves a simple story’ (2002:323).

4.iv.6.b: Governing Principles for Compression

Continuing briefly with the governing principles, these particular guidelines are intended to maximize compression amongst the vital relations in a blend. They instantiate certain of the subgoals above, and are guidelines for best achieving compression. For example, a number of the principles (which will not be described in depth in this section) aim to identify those relations which have a scale which can be condensed (Time, Space, Change, Cause-Effect, Part-Whole, Property, Similarity and Intentionality) and those which cannot (Analogy, Disanalogy, Identity, Representation), and should instead be compressed into other relationships. Others encourage ‘borrowing’ into the blend those inner-space relations which have ‘existing tight coherence at the

human scale' (2002:324) over those which do not, and syncopating single relations by 'dropping out all but a few key elements' (*ibid*).

4.iv.6.c: Other Governing Principles

These complement (although also occasionally conflict with) the principle of compression. First is the **topology principle**, which 'resists Compression that eliminates important topology' (2002:328) and so ensures the blend contains relations in the blend which match the relations of their counterparts in other spaces; *topology* here means the mental representation of a scene in terms of image-schemas, force dynamics and vital representations (2002:104-5).³⁴ The **pattern-completion principle**, which aims to compete blends by using existing and already-integrated tightly-structured patterns if an additional input is required, rather than any alternative. The **integration principle** simply requires a blend to be integrated, so that the blended space can be manipulated and remembered as a single, coherent unit (and so this requirement leads to coherent selection from the inputs to form fused elements and thus a tightly-integrated blend). The **principles of vital relations** promote vital relations where these are necessary by context (blends which explore similarities should maximize vital relations within the blend and between its inputs) and aim to intensify the relations where possible, to be 'sharper, simpler and stronger' (2002:330). The **web principle** promotes the relationship between element links in the blended space and the input spaces, so that the blend should be able to be related back to the inputs easily. The **unpacking principle** states that blends should themselves be able to trigger the whole network; that

³⁴ 'The different topological properties of individual spaces [...] give rise naturally for different possibilities for matches between them. On the one hand, we can look at what is inside spaces and build correspondences between one space and another based on similar topologies. Thus, for instance, we might map a linear scale in one space onto a linear scale in another space, or a source-path-goal image-schema in one space onto a course-path-goal image-schema in another space, or the force-dynamic pattern of Caused Motion in one space onto a force-dynamic pattern of Caused Motion in another space. In all such cases, the vital relations of Identity or Analogy apply across spaces to the topology of scales, image-schemas and force-dynamic patterns inside mental spaces.' (2002:105). Fauconnier and Turner here bring in the cognitive-semantic work of Talmy (eg 2000) and others to conceptual integration, emphasising that the theory does not exist apart from the rest of cognitive linguistics.

a reader/hearer confronted with a blend should be able to work out its inputs (necessary for comprehension). Finally, the **relevance principle** fulfils general pragmatic demands that the elements in a blend should be relevant to the construction and interpretation of the blend and the larger network, and that important relations between the input spaces should have correlates with inner-space relations in the blend.

4.IV.7: REPRESENTATIONAL CONSIDERATIONS

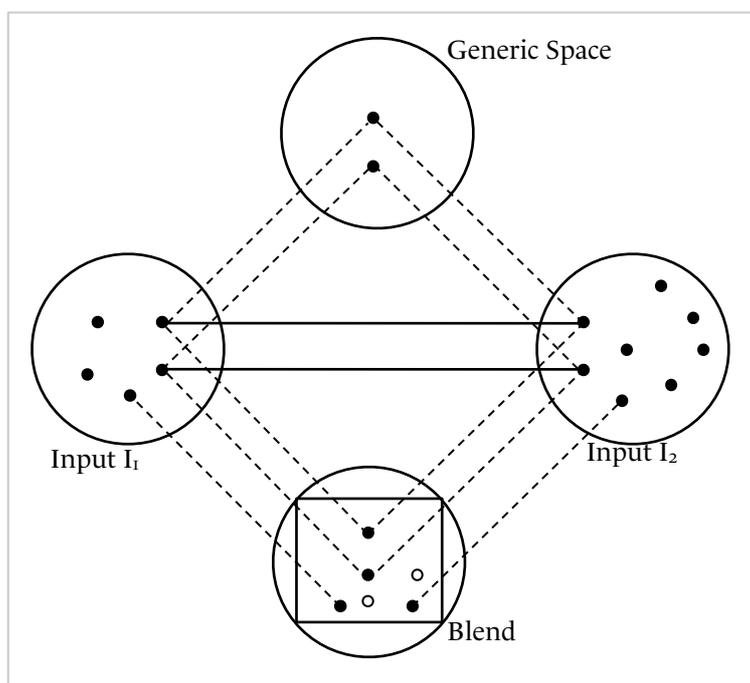


Figure II: A Schematic Outline of a Minimal³⁵ Conceptual Blend, from Fauconnier and Turner 2002:46.

Due to the length of the texts analysed in the following chapter, it is advisable to engage here with the representational standards of CIT analyses. Usually diagrammed visually (see Figure II above), there are certain representational requirements of CIT which are defined in Fauconnier and Turner (2002:40ff). Firstly, a circle is used to represent a mental space, and elements within each

³⁵ That is, containing only the minimum constituents required for a blend to take place; two input spaces and one blend, as opposed to more complex examples.

space are indicated by dots (or sometimes icons). Solid lines represent matching and counterpart connections between inputs, while broken lines indicate projections from inputs (to a blend or a generic space). A square within a circle represents ‘emergent structure’ and identifies those blended spaces in diagrams with multiple inputs and multiple blended spaces. Consequently, dots inside a blended space (a square inside a circle) with two broken projection lines represent fused elements, while those with only a single projection line are non-fused. ‘Empty’ dots are emergent entities (see the blend feature of completion above).

Other linguists and theorists make various changes to the Fauconnier and Turner diagrammatic guidelines.³⁶ For example, many writers (such as Mandelblit 2000, Coulson and Oakley 2003, Coulson 2001 and Ungerer and Schmid 2006) omit the generic space as ‘it does not add anything that we regard as essential for understanding the network model’ (Ungerer and Schmid 2006:259). More significantly, Coulson 2001, amongst others, often dispenses with the diagram and instead represents the process of integration using a tabular format, although in complex relationships she reintroduces a generic column in the representative table. This practice will be followed where possible in the analysis below, for the sake of space.

4.IV.8: CRITICISMS

And now there is *blending*, or *conceptual integration theory* (CIT), originally developed by Mark Turner and Gilles Fauconnier. Once again, some of us are in the fascination stage, while others are already savoring long-awaited answers. Still others think it is not enough of a theory yet and are cautiously awaiting the moment when blending becomes so rich theoretically that it will start creating problems for itself. But, at least to my knowledge, no-one has as yet called it a passing fad.

³⁶ It is worth nothing that Fauconnier and Turner themselves ignore their own rules, such as the highly non-standard diagram at 2002:62.

4.iv.8.a: A *Post Hoc* Theory

A very common – and generally valid – criticism of Conceptual Integration Theory is that it is a *post hoc* theory. An analyst looks at discourse, in this criticism, and realizes that there is probably a conceptual blend in the section under consideration *before* the analysis takes place. In this way, the theory does not predict the existence of blending but rather retrospectively describes the details of the blend, and nor can it make any claims to be theoretically free from intuition or bias on behalf of the analyst (see §4.iv.5.c on *Mappings* above).

Fauconnier and Turner do strongly address the problem of making falsifiable predictions (2002:54-5):

This objection is simply wrong. Theories of probability, subatomic particles, chaos, complex adaptive systems, evolution, immunology, and many others could not get off the ground as sciences if they were required to offer models in which the specified inputs determined unique outputs. [...] Given the nature of the mental operations of blending, it would be nonsense to predict that from two inputs a certain blend must result or that a specific blend must arise at such-and-such a place and time. Human beings do not think that way.

Although this quote veers close to neutralising the analytic abilities of CIT, Coulson and Oakley (2000:192) further clarify:

While post hoc analysis does indeed have [a danger of the fallacious imputing of cause], we note that most fruitful theories of meaning begin with this sort of analysis. [...] So, while we agree that blending analyses must move beyond post hoc analysis, we think that post hoc analyses of texts will necessarily play a basic role in building and refining models of online meaning construction.

Furthermore, these appear to be problems of psychology, not of linguistics. Particularly in the field of discourse analysis, the charge of operating *post hoc* is a common one (see, for example, Widdowson 2004:109 on Critical Discourse

Analysis), however some steps have been made towards moving away from the necessity of being *post hoc*, most notably with the use of digital methodologies in stylistics and discourse. As a computer cannot interpret discourse apart from in the ways in which it has been programmed, any attempt to integrate digital methodologies into discourse analysis tends to involve confronting certain assumptions of interpretation and systematising them wherever possible (eg the nature of semantic prosody, or the interpretative strength of lexical collocations). The present thesis, like the key CIT theorists above, recognizes with the *post hoc* nature of CIT and aims to provide some small steps towards moving away from this towards a more empirical technique.

While the central aim of the present thesis is a description of how strategies of reification operate in discourse *where such analogies have been proven to occur* (see previous chapters), and it therefore requires appropriate descriptive (and not predictive) techniques, there are ways in which the methodologies of establishing reification can be expanded and integrated into the CIT analysis, as §4.v below attempts to demonstrate. While the charge of *post hoc* operation could possibly be avoided for this thesis, the preferred alternative would be to go as far as possible with a clear discovery procedure in mind, rather than a descriptive procedure after the case. The preceding chapters can therefore be seen as the first stages of an empirically-based methodology for using CIT; first, as the saying goes, catch your domains.

Finally, it should be pointed out that the *post hoc* charge has not prevented many linguists from employing blending in fruitful and interesting ways (see, for example, Coulson 2001, Dancygier 2006, Dancygier forthcoming, the 2006 special issue of *Language and Literature*, the 2005 special issue of the *Journal of Pragmatics*, or the 2000 special issue of *Cognitive Linguistics*). Furthermore, as Fauconnier and Turner state in a quote above, the mind ‘may make many parallel attempts to find suitable projections, with only the accepted ones appearing in the final network’ (2002:72) – with such a claim made, standard

within network and probabilistic theories of meaning (see *inter alia* Bod *et al* 2003), a charge of the interpretations here being slippery and ad-hoc may nonetheless be countered by making explicit reference to the probabilistic psycholinguistic selection of a ‘successful’ blend amongst others created in parallel.

4.iv.8.b: A Broad Theory

The sheer breadth of the theory, taken alongside its subject domain of all of semiotic interpretation and conceptualization, is perhaps at odds with detailed linguistic analysis, at least at this stage in the development of the theory.

Fauconnier and Turner have even argued that the perception of motion is a result of conceptual blending:

Fauconnier and Turner argue that the experience of motion is the result of the visual system’s integration of two separate events into the unified percept of motion. In a conceptual integration network, one input space represents light 1 flashing in place p_1 at time t_1 , while the other input space represents light 2 flashing in place p_2 at time t_2 . In the blend, a single light, which we will call light 3 maps onto light 1 and light 2. By composing the events in the two inputs, light 3’s transition from p_1 at t_1 to p_2 at t_2 is understood and experienced as the result of motion.

(Coulson and Oakley 2000:182)

The criticism of overall breadth and, more particularly, too powerful a theory, is an early one (although occasionally repeated without reference to the changes in the theory over time). After early criticisms, Fauconnier and Turner introduced the governing principles outlined above to limit the theory. As Veale and O’Donoghue (2000:254) say:

Primarily, consider that theoretical utility is inversely proportional to expressive power, and that overly powerful theories have little cognitive status, since – scientifically – one should seek the least powerful theory that accounts for the most facts. Additionally, because conceptual integration expands the descriptive

options open to the cognitive theorist, there is a danger that some analyses will have the appearance of cognitive ‘just-so’ stories unless these analyses can be motivated in terms of independently compelling principles or constraints. For this purpose, Fauconnier and Turner provide a variety of structural principles that determine the optimality conditions under which integration can occur in a network, sufficiently reducing the theory’s degrees of freedom, and thus the arbitrariness of its analyses.

They conclude that the theory, when considered with these principles, is ‘sufficiently well articulated to support an algorithmic view of its component processes and representations, and sufficiently well constrained as to make this algorithmic view computationally tractable’ (2000:279), and this view is taken in the present thesis.

4.iv.8.c: Granularity

One issue perhaps arising from a discursal approach is that of granularity; that is, is the unit of analysis in blending a discourse, a cultural system, a sentence or even a word? CIT has been used for the description of neologisms (*caffeine headache* is a famous example), metaphors, advertisements, entire stories and whole systems of thought. (The accessibility of *The Way We Think* has also led many amateur cultural theorists, particularly on the Internet, to explain any number of matters of social semiotics in terms of blending.) Many discourse analysts are familiar with theories which operate in similar ways at various scales of grain, however some linguists (such as those influenced by Hallidayan systemic-functionalism where a rank scale system often divides various ‘layers’ of linguistic representation) object to this system as unspecific. This is, however, not the case in specifically cognitive models: cognitive scholars (such as Markman 1999) explicitly require that cognitive models *must* use representations at multiple levels of granularity (see also Gibbs 2000:352). This point explains in part the assertions above regarding the wide-ranging nature of CIT analyses, and the different types of analytical procedure that a CIT analysis

can follow. The present thesis, then, focuses primarily on words as entrypoints to semantic domains.

4.iv.8.d: Use

The final issue discussed here with regards to CIT's use is the perceptual issues which arise in the field in its textual use. In particular, in the field of cognitive linguistics, semantics, and discourse analysis, conceptual integration has frequently been used to explain strange, unique or complex types of linguistic expressions – literary metaphors, advertising language, neologisms and so on. This can be a barrier to using it as Fauconnier and Turner outline above, as 'a routine, workaday process' (1998:113). However, as Coulson (2001:115) explicitly states:

[although] blending is frequently employed for sophisticated feats of reasoning, its intermediate products are cognitive models whose plausibility spans the gamut from chimerical, to merely bizarre, to downright trite. [...] cognitive models built in blended spaces can yield productive inferences in spite of, and sometimes even *because of*, their strange properties.

While it is not argued here that popular science is a straightforward text-type (see §I.iii in the Introduction above), it is still not an overtly 'bizarre' discourse style, and so CIT's use in this thesis is viewed as justified.

4.V: REIFICATION ANALYSIS IN PRACTICE

Finally, this section describes the analytic procedure to be used in the following chapters and explicates the methodology and rationale behind it. It concludes with a brief proof-of-concept analysis of a reified abstract conceptualization, in preparation for the lengthier analysis which follows in Chapter 5. As described above, the practice of CIT analysis is a procedure which operates from identified mental spaces through a series of principles into integrated mental representations of meaning in discourse. In the process of finding these

integrated representations, which are one of the key results of this thesis, the major problem from the point of view of a researcher interested in using empirical and non-impressionistic methods in a CIT analysis is working against the *post hoc* charge outlined in §4.iv.8.a above. This complaint, which is a legitimate one, operates mostly with regards to input spaces, as the theory is relatively robust when it comes to the application of the various governing principles and vital relations in the creation of the blend. This thesis is not aimed at a wholesale revision of CIT, but it does claim that certain of the methods used in previous chapters point the way towards the use of lexicological data such as that contained in the *Historical Thesaurus* to assist with the creation of the blend.

In so doing, the following sections outline suggested improvements to the identification and construction of mental spaces, before moving on to a worked example of the proposed methodology.

4.V.I: THE IDENTIFICATION OF INPUT SPACES

Firstly, the primary requirement in an analysis within Conceptual Integration Theory is that the various input spaces (or semantic domains) which are activated by a text are described. Within CIT, these are inherited from Fauconnier's theory of mental spaces, where discursual creation of a particular space is signalled by a range of linguistic expressions he calls *space builders*, 'overt mechanisms which speakers can use to induce the hearer to set up a new mental space' (Fauconnier and Sweetser 1996:10). These generally include locatives (eg prepositional phrases, such as 'In Glasgow', 'In 2006', etc), adverb phrases ('really', 'theoretically'), and connectives (such as conditional if-then structures, or the either-or structure). He also identifies certain 'underlying subject-verb combinations', such as 'Max believes', 'Mary hopes', 'Gertrude claims', and so on (Fauconnier 1985:17), which appear to combine the major epistemic/factive, implicative, and boulomaic verbs of English. The major focus of much work in mental space theory, before Fauconnier moved onto

Conceptual Integration Theory, was in the description of logical counterfactuals (see, eg, most of Fauconnier 1985 and Fauconnier and Sweetser 1996), which explains the bias in this list towards those elements which introduce counterfactual constructions. No details are given as to the way in which this list was arrived at. This causes a problem for an empirically-based investigation of mental spaces, as is outlined below.

Fauconnier and Turner's own oft-cited example of 'That surgeon is a butcher', meaning the surgeon is a poor one, is said to instantiate two mental spaces, surgery and butchery, however there is no evidence of a space builder here – no locatives, adverbials, connectives or epistemic verb constructions are present, but is rather made up of two noun phrases linked by a copula, one acting as subject (introduced by a deictic demonstrative) and the other a predicate-nominative subject complement. The two nominal referents, *surgeon* and *butcher*, are elements within spaces, as Fauconnier uses the term, but there is no explicit sign that two spaces exist (as the Fauconnier and Turner example requires). The definition of an element in the Fauconnier system is that which is instantiated by a noun phrase (generally made up of naming elements, describing elements and pronominals; Fauconnier 1985, chapter 1). While the initial unbound nominal element in the surgeon-butcher example can be used to create a base space (as in Fauconnier and Sweetser 1996:10), this only justifies the creation of a surgeon base space, not a related butcher space. Compare the invented example 'John is married to Mary', where the copula straightforwardly creates a base space with two elements, John and Mary, and a property of a marriage relationship between them. (This is explicitly made clear in other examples used in Mental Space theory; Fauconnier's own example of a similar construction (1985:22) is 'Henry's girlfriend, Annette, is Swedish; but Len believes that she's his wife, that her name is Lisa, and that she's Spanish', which he then describes as having a base space with Henry, Annette, girlfriend of Henry, Swedish, etc, and a counterfactual space where these elements link to the alternatives that Len believes. The key point is that in Fauconnier's own

example, neither the subject, complement, nor the copula create a second mental space.) One may argue, of course, that the blended space which results from the sentence ‘That surgeon is a butcher’ is a single mental space in the mould of ‘John is married to Mary’, and therefore operates successfully within the theory, but then we face the logical conclusion that there exist mental space builders which operate to create input spaces to blends but not to create blended spaces or spaces not part of the conceptual integration process. This postulate would lead to a hierarchy of mental spaces with each having associated space builders of different types, and such a statement is not supported within the published framework, or within the general point in CIT that mental spaces are basic-level units, and not subject to further division. Fauconnier and Turner (2002:40ff) themselves define and describe the process of creating spaces entirely ostensively, though worked examples, rather than by defining them linguistically. Other work in CIT takes the process of space creation as a given, focusing mainly on the higher-level effects and patterns of integration. (This is perhaps because of the relative youth of the theory, although it should be noted that similar issues apply to the older Lakoff *et al* Conceptual Metaphor Theory, which is similarly underdeveloped at the mechanical level.)

4.v.I.a: Relevance Theory and Input Spaces

The postulate of a hierarchy of space builders may be useful in future research, but does not appear to be immediately helpful for the present thesis, and would result in theoretical questions which are beyond the scope of this work. Instead, I would prefer to suggest that each noun phrase has the potential to be a space-builder, and that a reader applies a pragmatic judgement to judge, in the process of meaning construction, if the new entity has a strong relevant link to the preceding discourse or if it is likely to constitute a departure from the text which justifies a new mental space. The framework I have chosen to site this within is that of Relevance Theory (RT), a cognitive approach to pragmatics in the Gricean mould which places the principle of relevance to be

paramount in the interpretation of discourse (Sperber and Wilson 1995). The reason for my decision is that Relevance Theory is already explicitly cognitive, and has been applied alongside existing cognitive theories of discourse fruitfully, while being described as fundamentally complementary (see, *inter alia*, Ungerer and Schmid 2006:288ff, Tendahl and Gibbs 2008 and Tendahl 2009). Despite these approaches, however, this thesis only uses RT's pragmatic approach to solve the problem of space selection sufficiently to allow data from other sources, such as the *Historical Thesaurus*, to be incorporated. RT's main assertion is that:

Roughly put, an input to a cognitive system is relevant when on the basis of existing information the input yields new cognitive effects. Cognitive effects are achieved when a speaker's utterance strengthens or contradicts an existing assumption or by combining an existing assumption with new information to yield some new cognitive implications. (Tendahl and Gibbs 2008:1831)

The applicability here, to solve the problem of when to identify a referring expression as a new space, is clear; new input spaces are those which do not have relevance to previous input spaces, which is to say, referring expressions which fulfil the criteria of strengthening, contradicting or adding to a previously-identified input space (in a very strict lexical interpretation) do not create a new input space, but are rather additional elements which have an anaphoric (or, in rarer cases, cataphoric; see §5.ii.1.b below) relation to an existing space. That which remains can be justifiably claimed to be a new input.

A general outline of how RT works in the interpretation of discourse is below, adapted from Yus (2009):

1. Given any sentence, there are a lot of meanings which could be 'decoded' from that sentence – any given sentence has a wide range of different interpretations, both literally and with regards to what the speaker is trying to imply.
2. All these different interpretations are not equal; in particular, they are graded in terms of their *accessibility*. This is because human beings can cognitively deal

with a wide range of situations, but some situations require more mental processing than others. [Very similar claims are made by other cognitive linguists under various terms such as *processibility*.]

3. Hearers therefore rely on a powerful criterion when selecting the most appropriate interpretation from this wide range. This is the assumption that a speaker is being maximally relevant.

4. This criterion then makes it possible to select one particular interpretation from that wide range of possible interpretations – and as speakers and hearers are geared towards mental efficiency, then when a hearer goes through this accessibility-ranked list of interpretations, they will stop when they come to the first candidate which the criterion tells them must match their intended interpretation.

This procedure is guided by the following definitions and principles, adapted from Sperber and Wilson 1995 and 2005, and Wilson and Sperber 2002 and 2004:

a. Other things being equal, the greater the positive cognitive effects achieved by processing an input, the greater the relevance of the input to the individual at that time. As above, positive cognitive effect is a worthwhile difference to the individual's representation of the world: a true conclusion, for example, or the strengthening, revision, or abandonment of available assumptions.

b. Other things being equal, the greater the processing effort expended, the lower the relevance of the input to the individual at that time.

c. The ostensive stimulus is relevant enough to be worth the addressee's effort to process it.

d. The ostensive stimulus is the most relevant one compatible with the speaker's abilities and preferences.

Therefore, should a speaker be engaged in comprehending 'That surgeon is a butcher', as above, they will first interpret the subject 'That surgeon' as an element within the base mental space, and then encounter the subject

complement 'a butcher'. The possible interpretations are literal and figurative; the surgeon also works as a butcher, or the surgeon is figuratively a butcher. (The assumption is made here that for the purpose of this thesis, and as established in previous chapters, the texts to be examined here will have a range of figurative meanings. Other options will be available for other possible sentences; for example, 'That surgeon is his mother' will have a range of interpretations due to the polysemy of *mother*; see Lakoff's (1987:74ff) discussion of this multiplicity.) These interpretations require either an expansion of the base mental space or the creation of a new one, respectively. For analysis, the interpretations can be ordered by processibility/accessibility with the literal interpretation first, as a literal interpretation would axiomatically be easier to interpret than the output of a process of blending. Contextual cues will, however, promote alternative interpretations (this sentence, when said with the intent to state a judgment on the skills of the surgeon, would often be said with stress on *butcher*, and probably with a rise-fall intonation on those syllables to signal disapproval; there may be further discursual cues to this interpretation). In this instance, the greatest positive cognitive effect to be created in the mind of the hearer is important information about the skill of the surgeon, rather than just additional information about the surgeon's second career (see (a) above), which increases the likelihood that this is the most relevant interpretation, given assumption (c) above, alongside step 4 in the interpretative procedure. The probability is therefore that the discourse interpretation is figurative, necessitating a new mental space to be created by the second noun phrase.

4.v.I.b: Relevance Theory Principles for Input Spaces

An adaptation of the principles above for the purpose of input space identification is therefore given below. Note that the default assumption throughout is that any new referring expression is a new mental space unless the RT guidance indicates otherwise. This will result in a maximally-large number of input spaces, rather more than impressionistic analysis would give

of the same text, and the analysis below shows ways in which these are compressed where possible.

- a. Other things being equal, the greater the positive cognitive effects achieved by processing a referring expression, the more likely it is that this expression is not a new input space but relates to one previously enacted. The relationship between further elements and referring expressions to positive cognitive effects is most clear in the case of elements which strengthen, revise, or alter previously-enacted spaces.
- b. Other things being equal, the greater the processing effort expended on a referring expression, the more likely it is that this expression enacts a new mental space than adds to a previously-enacted space.
- c. Any referring expression in a discourse is relevant enough to the purpose of the discourse to be worth the addressee's effort to process it. Therefore, any referring expression must either relate to an existing space or create a new one.
- d. The interpretation of a referring expression (that is, as a new input space, as a reference to a prior input space, or as an element in an existing space) which is most compatible with the speaker's abilities and preferences is, other things being equal, the most likely interpretation.

Principle (d) here is perhaps overly generic, and is included here mainly for the sake of completeness. Also, principle (c) is the principle which supports the view above that the default position is that referring expressions create a new input space if identification as an previous-space element is problematic. It is preferred to follow this principle of proliferation than to be overly spontaneous in the analysis of a text. The remainder of this list highlights two core principles; that positive cognitive effect promotes a referring expression's likelihood of being a previous-space enactor, and that processual complexity promotes the likelihood of a referring expression being an enactor of a new input space. An analyst therefore needs to balance these likelihoods to arrive at a description of the referring expressions concerned. However, interpreting this

balance, based on established principles, is preferable to an impressionistic best-guess analysis.

Note that the use of RT here *for input space selection* is different from the use of RT within the more general study of figurative language, which has been examined by other researchers (see above). In many ways, the use of RT in other parts of the blending process is rendered functionally unnecessary by the existence of the relevance principle in CIT, at least for the purposes of the present thesis. It is employed here simply as a consistent mechanism to judge the likelihood of the necessity for a further mental space. This use of RT is preferable to the alternative above (of a hierarchy of space builders), in part because it links to Fauconnier and Turner's statement of the importance of relevance in their relevance principle (2002:333ff). It is also an instance of the present thesis' preference for a small, theory-appropriate supplement to the existing CIT apparatus rather than a overhaul of a core part of the theory, given limitations of space and scope. It is notable that RT researchers would claim that the theory would be able to handle the interpretation of the full sentence, although this is by no means a certain claim, and nor does it mean that analysis in this fashion would end up being descriptively adequate for the purposes of discursal research (see Tendahl 2009 for details). The implementation of RT within the analytic procedure of these chapters is therefore simply as a low-level pragmatic mechanism with regards to the identification of input spaces to the conceptual blends the procedure aims to demonstrate; the CIT procedure is considered authoritative, but when it comes to identifying which elements indicate a mental space which is not introduced by an explicit counterfactual world-builder, the RT system is used as a backup solely to determine the necessity of a new mental space.

4.V.2: THE CONSTITUTION OF INPUT SPACES

When a space is identified, a further common problem with CIT is that the spaces are named in a somewhat impressionistic way. Often they are given the

names of their major element (eg SURGEON and BUTCHER), although occasionally they are not, depending on whether or not the analyst is naming them with an eye to their future use in other blends. Therefore, a straightforward interpretation of ‘He’s very sweet’ could involve instantiation of any one of a number of relationships (to use conceptual metaphor nominalizations, this could be any of BEING SACCHARINE IS BEING ENDEARING, SWEETNESS IS ATTRACTIVENESS, PLEASANT TASTE IS POSITIVE EMOTION, OR TASTE IS EMOTION). The naming and internal makeup of the input domains are subject to any preference of the analyst, and are rarely if ever systematized. This applies to all of cognitive linguistics, and is particularly apparent pedagogically, where any teacher can be struck by huge variability in student answers to requests to formalize domain relationships within any given example. This variability is convenient, but can on occasion serve to give an impression of a *fait accompli* in the analysis – in other words, an analyst selects the space name and the space elements of the analysis, often giving those which fit neatly with their conclusions.

This is not very surprising, given the lack of a comprehensive list of semantic domains, or even agreement on what makes a semantic domain. However, with the completion of the *Historical Thesaurus* there is a strong case to be made that semantic domains and mental spaces can be accurately mirrored in *HT* categories. While there is an argument that *HT* domains can be used as a lexicographical shorthand to refer to a domain, there is also a deeper argument that their structure reflects more broadly the cognitive construal of categories in English. While the *HT* itself makes no claim to represent mental spaces, its editors do state:

Although the project was started before the current cognitive semantics paradigm became dominant, that paradigm has retrospectively proved sympathetic to the problems involved in categorizing large quantities of lexical data. The development of prototype theory, which allows for fuzzy sets containing both good and less good examples of the central concept, challenges

the either/or basis of Aristotelian category assignment and liberates semanticists from a narrow notion of synonymy as an organizing principle. *HTOED*'s³⁷ synonym groupings are prototypical in nature, with a clear core of obvious members shading off into the less obvious, and ultimately into cognate categories. (Kay *et al* 2009: xix)

The various pieces of research done within cognitive linguistics using *HT* materials (eg Allan 2009; for more, see the Selected Bibliography at pp. xxxii-xxxiv of Kay *et al* 2009) supports the assertion that the *HT* has a strong claim to being a representation of semantic categories as they are instantiated in lexis. The data-driven approach means that, based on citation evidence, *HT* categories were constructed according to prototype theory, even when prototype theory had not yet been fully described theoretically. In this way, any mental space activated in a text can be assigned an *HT* category reference, giving it an independent label and allowing the domain to thus be named for analysis in a way which is less impressionistic than the default CIT system.

In the absence of a semantic tagger using the *HT* (see §2.iii above, and §6.iii below), the process can be crudely undertaken as follows: a wordlist of the relevant extracts to be analysed is created in the same way as the Ordered List version of the full texts (see §2.ii.3 above), giving a simple list of lemmas used in the text, filtered for stopwords, alongside a number representing where it comes in the stopword-filtered text. Each lemma is then linked to the *HT* to give a list of all the *HT* domains for each lemma. This list is then summarized by restricting the lemma's category number to the fourth level of the *HT* (the fourth hierarchical level down from the root of the Thesaurus, containing such basic concepts as the *Sun*, *Weather*, *Age*, *Sex*, *Safety*, *Direction*, *Hope*, *Dislike*, *the Army*, *Law*, etc) and counting how often those possible categories for each lemma appear at each level. For example, the lemma *speak* appears 49 times in

³⁷ As outlined in §2.ii.2.b, *HTOED* is the abbreviation for the *Historical Thesaurus of the Oxford English Dictionary*, the name of the print version of what this thesis simply calls the *Historical Thesaurus*.

the *HT*, and 11 of those are within the wider category *02.08.03 Speech/act of speaking*, giving that category a ranking of 11, with *02.08.03.07 Conversation* having a rank of 4, *02.08.06 Statement/declaration* 3, and so on down to rarer uses like *03.08.04 Indication* and *02.08.03.03 Narration*, which contain only one *speak* lemma each. Distilling the number of possible categories by higher-level category is a useful measure to decide which meaning of the lemma could be used in a future tagger. An analyst can therefore use this list to manually tag each new mental space with the appropriate category, with the assumption that it would likely be the category containing the most realizations of that lemma that would be the default category to assign. (This is a matter of convenience rather than a methodological requirement.) This is a very rough measure, given that some words already in higher-level categories will appear alongside the truncated category references, meaning the analyst has to take these into account. On occasion, there will also be a failure of lemmatization, necessitating a manual lookup of the *HT*. Nonetheless, it is a useful, although imperfect, tool for the analyst to have in their possession for this stage of the process.

4.V.3: THE INTERNAL MAKEUP OF INPUT SPACES

Finally, once input spaces have been identified and labelled with a level four *HT* category, their internal structure is the last constituent of the input space before the various CIT analytic procedures are put into action. This is particularly difficult to systematize, as it draws upon schema knowledge beyond the text, using a reader's generic knowledge to fill in any gaps in the elements represented in a space beyond that which is explicit in the text (see Schank and Abelson 1977 for an authoritative description, and, *inter alia*, Emmott and Alexander 2009/2010 or Stockwell 2006 for an overview). These gap-filling elements in the mental space are almost by definition going to be chosen *post hoc* by the analyst, as otherwise any analysis would turn into a

lengthy list of attributes, relationships and encyclopaedic knowledge on any given topic.

A proposal to add some sense of extra-analytic input to the procedure is therefore to prefer to use non-space-building referring expressions within the text itself as elements, and wherever possible to reinforce the presence of additional schema-filled elements with reference to external information sources (eg dictionaries and encyclopaedias; a further preference for advanced learner's dictionaries could be recognized). This is perhaps a relatively low-priority requirement, though, given the nature of the problem of the blend.

4.v.4: AN ANALYTIC EXAMPLE

This section has described, in §4.iii and §4.v.1-3, the major features of the analytic procedure proposed for this thesis. This chapter therefore ends with an example of this procedure, as an illustration of the methodology in practice.

4.v.4.a: Extract Selection

In order to select a very short text for analysis, the five highest spikes for the two texts with the best systematic results in §4.ii.2 above (*FC* and *MP*) were manually read and a paragraph selected from the text in these spikes. No claim is made as to the nature or status of this paragraph, other than it is not in an extract chosen in §4.ii above, and it is a single paragraph with some substantial analogical content. Due to its brevity, it requires, however, some additional information from elsewhere in the text, as outlined following the extract below.

The main part of the text, then, is an extract from *FC*, a book concerned with explicating a model of quantum physics; the author explains the nature of spacetime in a string-theoretical universe:

Since we speak of the 'fabric' of spacetime, the suggestion goes, maybe spacetime is stitched out of strings much as a shirt is stitched out of thread. That is, much as joining numerous threads together in an appropriate pattern produces a shirt's

fabric, maybe joining numerous strings together in an appropriate pattern produces what we commonly call spacetime's fabric. Matter, like you and me, would then amount to additional agglomerations of vibrating strings.

(FC: Greene 2004:486-7)

For the sake of clarity, and to aid comprehension, the following is Greene's initial conceptualization of a string:

[...] every particle is composed of a tiny filament of energy, some hundred billion billion times smaller than a single atomic nucleus [...] which is shaped like a little string. And just as a violin string can vibrate in different patterns, each of which produces a different musical tone, the filaments of superstring theory can also vibrate in different patterns.

(FC: Greene 2004:17-18)

Combining these two short paragraphs, giving a total of five sentences, there is a sufficiently complex explanatory conceptualization that will suffice as an example of the data-driven approach to conceptual blending that is taken in this thesis. While the quotes above included ellipses to indicate the removal of extraneous material, below they are represented instead as contiguous text, and spaces referred to as if they are contiguous.³⁸

4.v.4.b: Identifying Input Spaces

The first stage is to identify the mental spaces created, which are here represented in **bold**, with explicit references back to these in *italics* and extra elements in previously-identified mental spaces underlined:

Since we speak of the '**fabric**' of *spacetime*, the suggestion goes, maybe *spacetime* is stitched out of **strings** much as a **shirt** is stitched out of **thread**. That is, much

³⁸ Questions of memory and the psycholinguistic retention of mental spaces during the process of reading are not here addressed (although see further Sanford and Garrod 1981). While this is a very interesting topic, it is outwith the remit of this present thesis. The contiguous version used here is treated as an entire stretch text, on the simple grounds that if it can be interpreted by a reader as a single text, it can be analysed as such. The space constraints for this initial analysis do not permit a longer text, resulting in the need for this short amalgamated text.

as joining numerous *threads* together in an appropriate pattern produces a *shirt's fabric*, maybe joining numerous *strings* together in an appropriate pattern produces what we commonly call *spacetime's fabric*. **Matter**, like you and me, would then amount to additional agglomerations of vibrating *strings*. Every **particle** is composed of a tiny filament of energy, some hundred billion billion times smaller than a single atomic nucleus which is shaped like a little *string*. And just as a **violin string** can vibrate in different *patterns*, each of which produces a different musical tone, the *filaments* of superstring theory can also vibrate in different *patterns*.

The mental spaces are identified according to the Relevance Theory guidance in §4.v.i.b above. As stated there, this might result in a maximal set of mental spaces (that is, a set which is as large as possible, and could be compressed for the sake of analysis), but as outlined in the RT principles, this is a virtue rather than a flaw in the approach. Later parts of this analysis describe criteria for the combination of input spaces.

To begin with, there are eight input spaces in the text under analysis. The RT principles above were used to decide which referring expressions created a new space; so that the NP 'the suggestion' was recognized as a metadiscoursal comment, rather than an NP requiring a new space, while 'maybe' here acts as a hypothesis indicator, rather than a counterfactual enactor. Metonymic relations (such as the material-object metonymy *shirt* and *fabric*) were recognized as different mental spaces due to a reluctance not to collapse metonymic relationships at this stage.³⁹ The RT guidance here is that the verb

³⁹ This relates to the discussion in §1.ii of the complex nature of metaphor and related phenomena, including metonymy. For now, the main issue is whether the metonymic relationships in this text are intentional or accidental; that is, is there an important and principled distinction made between shirts and fabric in the discourse, or would this distinction be necessary for its interpretation? On balance, as a general point, it is best to follow the principle above of positing extra input spaces rather than fewer. An alternative would be a minimal approach to metonymy where part-whole relationships are always inner-space relations by default until the part is 'promoted' to full space status by virtue of having its own elements attached. This approach has its virtues, and perhaps might even have some psychological plausibility, but in the interests of reducing complexity and maintaining consistency with the maximal space approach outlined above, is not followed.

stitched conveys an active material-object relationship, meaning that the presumption of optimal relevance states such an active relationship would likely be made explicit due to an important textual distinction being necessary between the material and the object, rather than simply metonymy due to stylistic variation (the rest of the passage confirms this). Devices such as ‘since we speak of’ and ‘what we commonly call’ were ignored as their relevance was minimal and their use a purely rhetorical one of the generic inclusive *we*, rather than as pronouns functioning as legitimate space creators. Similarly, ‘like you and me’ is unlikely to be a new space builder, but rather functions as a highly-relevant (that is, immediate) example of matter. The remainder of the NPs in the extract are either space builders or elements in a previously-created space, depending on their discursal relevance.

4.v.4.c: Naming and Populating Input Spaces

The second stage of this analysis is to name and separate these spaces, and to do this the full text was put into the corpus procedure described in §4.v.2 above, giving a table with all the bare lemmas used in the short text alongside all the possible *HT* L4 categories they instantiate.

This can then be used to create an annotated version of the text, as below, this time ignoring anaphoric reference and extra elements:

Since we speak of the ‘**fabric**’ [01020914: *Textile fabric/an article of textile fabric*] of **spacetime** [010507: *Space*], the suggestion goes, maybe spacetime is stitched out of **strings** [01040705: *Atomic physics*] much as a **shirt** [01021002: *Types/styles of clothing*] is stitched out of **thread** [01020912: *Thread/yarn*]. That is, much as joining numerous threads together in an appropriate pattern produces a shirt’s fabric, maybe joining numerous strings together in an appropriate pattern produces what we commonly call spacetime’s fabric. **Matter** [01050104: *Materiality*], like you and me, would then amount to additional agglomerations of vibrating strings. Every **particle** [01040223: *Atomic chemistry*] is composed of a tiny filament of energy, some hundred billion billion times smaller than a single atomic nucleus which is shaped like a little string. And just as a **violin string**

[03110301: *Music*] can vibrate in different patterns, each of which produces a different musical tone, the filaments of superstring theory can also vibrate in different patterns.

Here we have a list of input spaces, contextually identified with a *HT* category. Many of these have more precise lower-level categories (eg 03110301080206|0701: *Strings of a bowable instrument* rather than 03110301: *Music*), but the higher-level categories are preferred for labels to avoid artificially narrowing the reference of the input space (the flexibility of which is emphasized in the various forms of mental space and CIT analyses).⁴⁰

The mental spaces activated in this text, with their first enactors and the later textual referring expressions which constitute elements in each space, are therefore:

fabric 01020914: *Textile fabric/an article of textile fabric*; stitched, pattern of thread

spacetime 010507: *Space*; stitched

strings 01040705: *Atomic physics*; stitched, energy, joined in pattern to form spacetime's fabric, vibrates, tiny string-like filament of energy, can vibrate in different patterns

shirt 01021002: *Types/styles of clothing*; stitched, has fabric

thread 01020912: *Thread/yarn*; stitched, joined in pattern to form fabric

matter 01050104: *Materiality*; agglomerations of strings

particle 01040223: *Atomic chemistry*; composed of strings

violin string 03110301: *Music*; can vibrate in different patterns, different vibrations produce different tones

⁴⁰ Note also that, for technical reasons, the *HT* category has no separating periods as in the original and in Part I above (010507: *Space* rather than 01.05.07: *Space*); these are functionally identical, but make it easier to link to a separate table of level 4 *HT* headings, edited for accessibility (as *HT* headings differ depending on the part of speech, choosing a single L4 heading for the entire category requires manual intervention).

With these mental space outlines, the third stage of analysis can proceed to the penultimate stage suggested as essential for a description of meaning construction through CIT.

4.v.4.d: Space Relationships

Each of the input spaces above will be given an identifier (the letter I, representing an input space, and a subscript number representing the order of input spaces, along with a title made up of an abbreviation of the *HT* category title, with part of its first enactor added for clarification where necessary) and placed into a table in order of their activation in the paragraph. In CIT Table 1⁴¹ below, therefore, the lexical referring expressions are grouped according to their similarity, on lexical grounds (either the use of the same lexeme (eg *stitched* or *vibrate*) or lexical-semantic adjacency (eg *thread/fabric*, textually supported by *thread* being defined as that which ‘produce[s] a shirt’s fabric’).

An additional piece of information which assists with the construction of a blending analysis is a list of prepositional and verbal relationships between these input spaces. It is expected that the groupings of referring expressions in the table rows above will describe a textual relationship between the corresponding input spaces. For example, the initial sentence of the text begins ‘Since we speak of the ‘fabric’ of spacetime’, giving an ‘I₁ of I₂’ relationship, just as ‘a shirt is stitched out of thread’ gives an ‘I₄ out of I₅’ relationship. Both of these are supported by the existence of a first row of shared vocabulary between the two spaces. The full list (with relationships listed with an identifier beginning with the letter R), is:

R₁: I₁ of I₂ (Since we speak of the ‘fabric’ of spacetime)

R₂: I₂ stitched out of I₃ (spacetime is stitched out of strings)

R₃: I₄ stitched out of I₅ (shirt is stitched out of thread)

⁴¹ The tables in this and the following chapter are labelled CIT Tables, to distinguish them from other tables elsewhere in the thesis. Many of them are of sufficient size and complexity to require them to be placed on a separate page.

I₁ – Textile fabric	I₂ – Space(time)	I₃ – Strings (atomic physics)	I₄ – Shirt (clothing)	I₅ – Thread	I₆ – Matter	I₇ – Particle	I₈ – Violin string (music)
stitched	stitched	stitched	stitched	stitched			
pattern of thread		joined in pattern to form spacetime's fabric	has fabric	joined in pattern to form fabric		composed of strings	
		vibrates can vibrate in different patterns					can vibrate in different patterns different vibrations produce different tones
		energy tiny string-like filament of energy					
					agglomerations of strings		

CIT Table 1: Strings Sample Extract, Input Spaces

R₄: I₅ join[ed...] in an appropriate pattern produces I₄ and I₁ (joining numerous threads together in an appropriate pattern produces a shirt's fabric)

R₅: I₄ [genitive] I₁ (a shirt's fabric)

R₆: I₃ join[ed...] in an appropriate pattern produces I₂ and I₁ (joining numerous strings together in an appropriate pattern produces what we commonly call spacetime's fabric)

R₇: I₂ [genitive] I₁ (spacetime's fabric)

R₈: I₆ consists of I₃ (Matter [...] would then amount to additional agglomerations of vibrating strings)

R₉: I₇ composed of I₃ (Every particle is composed of a tiny filament of energy [...] shaped like a little string)

R₁₀: I₈ operates just as I₃ (And just as a violin string can vibrate in different patterns [...] the filaments of superstring theory can also vibrate in different patterns)

Much as, in the expression 'spacetime stitched out of strings much as a shirt is stitched out of thread', is a preposition premodified by an adverbial particle of degree and so seems as if it should be included in this list; however as it links two processes rather than two referring expressions it is better placed in the next list. The relationships above relate to mental spaces, not groupings thereof. However, as some of the relationships above operate in parallel, it is also necessary to note those further relationships between them, denoted either by parallel constructions or prepositional/verbal relationships between mental space relationships rather than spaces. Identified by R_M followed by a letter (M here standing for *meta*, as these are technically metarelations), and with square brackets identifying parallel constructions where there is no prepositional/verbal relationship, these are:

R_{M1}: R₁ [duplicates] R₇ (I₁ of I₂/I₂ [genitive] I₁)

R_{M2}: R₂ much as R₃ (I₂ stitched out of I₃/I₄ stitched out of I₅)

R_{M3} : R_4 much as R_6 (I_5 join[ed...] in an appropriate pattern produces I_4 and I_1/I_3 join[ed...] in an appropriate pattern produces I_2 and I_1)

R_{M4} : R_5 [parallels] R_7 (I_4 [genitive] I_1/I_2 [genitive] I_1)

Note the difference in terminology above between *duplicates* (an identical relationship holds between identical input spaces) and *parallels* (an identical relationship holds between different input spaces). As could be predicted, these relationships mainly hold between closely-related spaces which share the lexical referring expression *stitched*.

This completes the list of information which would allow an analyst to begin a blending analysis.

4.v.4.e: Concept Integration

To summarize, thus far a list of mental spaces activated by the text has been listed and outlined, using RT as a guide to decide which constitute mental spaces and which constitute elements, with details taken from the *HT* database, and finally syntactically-signalled relationships between these spaces have been listed. This constitutes a detailed discovery procedure for conceptual blending, with as little native-speaker intuition or analyst intervention as possible, excepting that required to implement the guidelines of RT or to interpret syntactic relationships. This is in opposition to traditional conceptual integration analyses, where researcher instinct takes precedence over systematic and careful textual analysis. The final stage, the construction of mappings necessary to construct blended spaces, will similarly be guided by the constitutive and governing principles of CIT.

Firstly, each space is linked to others using the relationships list above. Each relationship is then labelled with a CIT Vital Relation (the principle that all mappings between spaces should instantiate a Vital Relation means that any relationship without a Vital Relation needs to be questioned) and finally, the overall process of blended space creation is then reinforced with regards to

rationalising the putative blending operations with the CIT governing principles. These blends will be identified in multiple stages, with first-stage blends being those which arise from the input spaces and relationships, whereas later blends may arise from those first-order blends themselves, in order to achieve CIT's aim of creating a single blend description for a complex concept.

4.v.4.e.i: Relationship Mapping

Diagrammatically, the relationships between input spaces are shown in Figure 12 below.

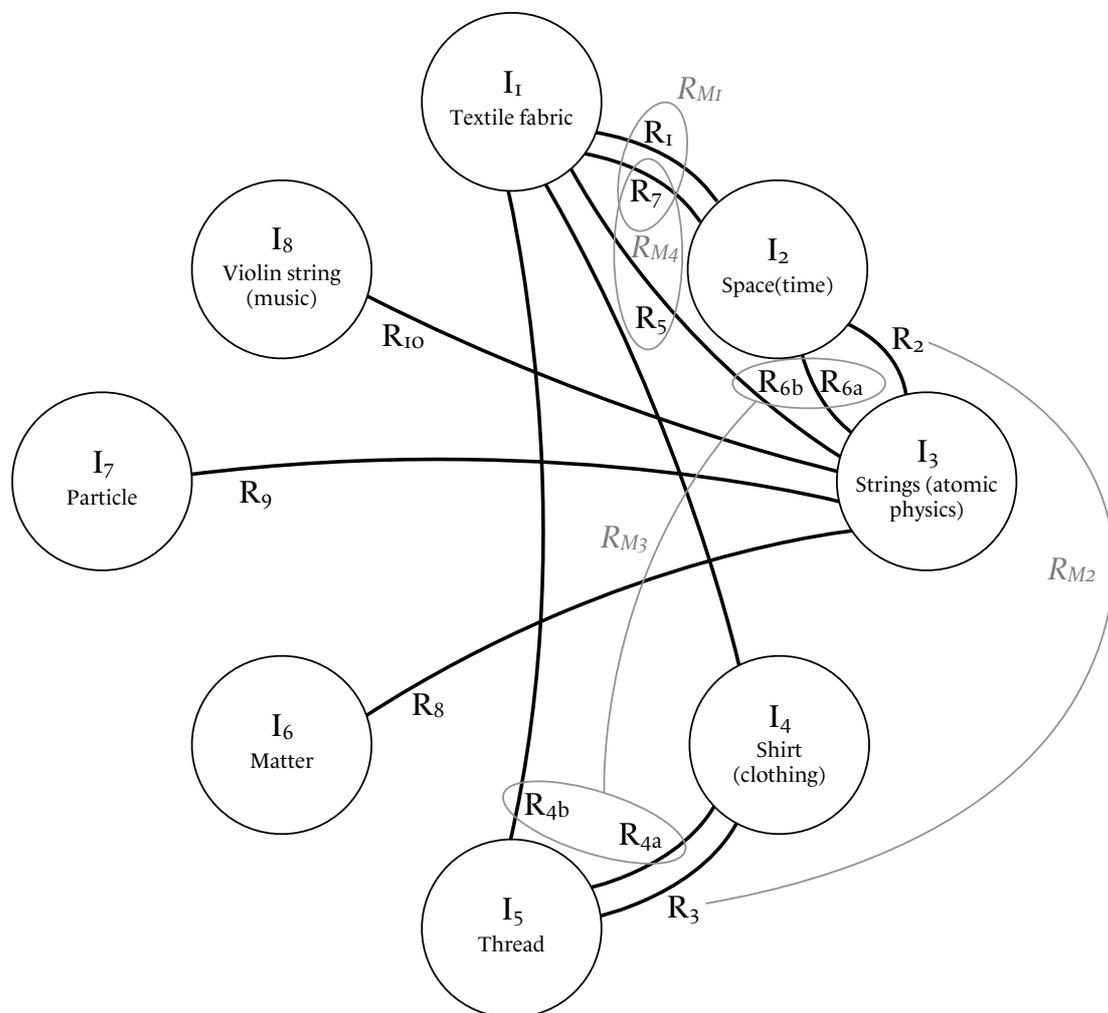


Figure 12: Strings Sample Extract, Input Space Relationships Diagram

Note that the relationships in grey are metarelationships (R_M -style) and that relationships with dual mappings (ie, R_4 and R_6) are denoted by subscript a and b to separate the two mappings. CIT Table 2 lists the type of mapping of each first-order relationship (Appendix II.1 summarizes and outlines the Vital Relations of CIT). Note that CIT Table 2 continues over three pages.

Relationship	Input Spaces	Vital Relation	Comments
R_1	I_1 of I_2 (Since we speak of the 'fabric' of spacetime)	Property	This profiles the nature of spacetime as an entity with a fabric-like nature. The <i>the x of y</i> construction often appears to indicate property (<i>the blueness of the cup</i> , to adapt one of Fauconnier and Turner's own examples in 2002:99-100, or <i>the beauty of the music</i> , etc).
R_2	I_2 stitched out of I_3 (spacetime is stitched out of strings)	Part-Whole and Cause-Effect	The double relation here and below refers to the use of the past tense fabrication verb <i>stitched</i> , which simultaneously profiles the assemblage of the entity and its constitutive parts. The Part-Whole relationship is therefore emphasized by the explicit reference to the parts which make up the whole, while the Cause-Effect relationship is noted by the use of the material verb which causes the parts to form the whole.
R_3	I_4 stitched out of I_5 (shirt is stitched out of thread)	Part-Whole and Cause-Effect	
R_4	I_5 join[ed...] in an appropriate pattern produces I_4 and I_1 (joining numerous threads together in an appropriate pattern produces a shirt's fabric)	Part-Whole and Cause-Effect	This is similar to R_2 and R_3 above, but with the fabrication verb being <i>join</i> .

Relationship	Input Spaces	Vital Relation	Comments
R ₅	I ₄ [genitive] I ₁ (a shirt's fabric)	Part-Whole	This could be considered ambiguous between Part-Whole and Property; however, while the above Property relation for 'we speak of the 'fabric' of spacetime' profiles the property of spacetime as being fabric-like, the construction 'produces a shirt's fabric' profiles instead, in Fauconnier and Turner's words, the shirt's 'most salient part' (2002:97). There is no other alternative for a statement profiling the material of an object other than Part-Whole or Property. While R ₁ has a very strong case for the 'fabric-ness' of spacetime being a property of that entity, there is no similar case that can be made here for a shirt.
R ₆	I ₃ join[ed...] in an appropriate pattern produces I ₂ and I ₁ (joining numerous strings together in an appropriate pattern produces what we commonly call spacetime's fabric)	Part-Whole and Cause-Effect	
R ₇	I ₂ [genitive] I ₁ (spacetime's fabric)	Part-Whole	
R ₈	I ₆ consists of I ₃ (Matter [...] would then amount to additional agglomerations of vibrating strings)	Part-Whole	Here, conversely to the <i>fabric</i> relationships, there is no profiling of the process of the assemblage of <i>matter</i> (that is, no explicit fabrication verb). Instead, the relationship is a simpler Part-Whole one between a constituent and an entity.
R ₉	I ₇ composed of I ₃ (Every particle is composed of a tiny filament of energy [...] shaped like a little string)	Part-Whole	

Relationship	Input Spaces	Vital Relation	Comments
R ₁₀	I ₈ operates just as I ₃ (And just as a violin string can vibrate in different patterns [...] the filaments of superstring theory can also vibrate in different patterns)	Analogy	These two vital relations relate to shared properties and roles. A comparison is drawn between the violin string and the superstring, and an Analogy relationship is formed with regards to the identity of the roles and actions of both.

CIT Table 2: Strings Sample Extract, Relationships

It is clear from these relations that the overall blend being created in this example uses a predominance of part-whole vital relations. While, to the best of my knowledge, no study has been published of vital relation patterning as it relates to discourse function, it is nonetheless likely that a part-whole preponderance here is a result of the use of prepositions to indicate particular CIT relationships. It is therefore likely that a number of these input spaces would have been compressed together in an impressionistic native-speaker-intuition analysis, although this is not supported by a direct text-based analysis.

Following the above outline of first-order relationships, the second-order metarelations are as follows in CIT Table 3:

Relationship	Input Spaces	Vital Relation	Comments
R _{M1}	R ₁ [parallels] R ₇ (I ₁ of I ₂ /I ₂ [genitive] I ₁)	Identity	This construction is identical in both R ₁ and R ₇ , with the exception of the choice of of-genitive versus the inflected genitive. This, along with the fact they join identical spaces, seems to indicate they should be the same relationship, rather than be considered separate ones.

Relationship	Input Spaces	Vital Relation	Comments
R _{M2}	R ₂ much as R ₃ (I ₂ stitched out of I ₃ / I ₄ stitched out of I ₅)	Analogy	Analogy requires ‘role-value’ compression, according to Fauconnier & Turner (2002:98). Here, the role is similar but the value differs (that is, they have similar operations and roles, but they are different entities).
R _{M3}	R ₄ much as R ₆ (I ₅ join[ed...] in an appropriate pattern produces I ₄ and I ₁ /I ₃ join[ed...] in an appropriate pattern produces I ₂ and I ₁)	Analogy	
R _{M4}	R ₅ [parallels] R ₇ (I ₄ [genitive] I ₁ /I ₂ [genitive] I ₁)	Analogy	Unlike R _{M1} above, with its Identity relationship, R ₅ and R ₇ relate to different spaces (I ₄ and I ₂ are the respective ‘possessors’ of fabric), and therefore an Identity relationship is not possible. Analogy is a more appropriate choice, given that the relationship is identical but the spaces are different, as above.

CIT Table 3: Strings Sample Extract, Metarelationships

At this stage, with as much information available as possible, two things can happen: firstly, the decision to combine disparate input mental spaces into a single space (given the second-order parallel and identical relationships outlined in the table above), and secondly, the description of the blending process itself can begin. As discussed in §4.iv.5.b above, the major aim of blending is to achieve compression over the relationships formed between input spaces. Therefore, these first- and second-order relationships in the tables above should be analysed for their potential to compress, and thus to form an integrated mental spaces blend. CIT Table 4 below does so, and begins with the second-order relationships, given that these relationships affect the first-order relationships which form them.

Relationship	Input Spaces	Original Vital Relation	Compressed Vital Relation/ Combined Space
R _{M1}	R ₁ [parallels] R ₇ (I ₁ of I ₂ /I ₂ [genitive] I ₁)	Identity	R ₁ should be combined with R ₇ , given the non-blended-space identity relation in this second-order relationship. (NB R ₇ in this table is therefore deleted.) The new R ₁ remains a property relationship between I ₁ and I ₂ .
R _{M2}	R ₂ much as R ₃ (I ₂ stitched out of I ₃ / I ₄ stitched out of I ₅)	Analogy	Analogy between two spaces creates Similarity in a blend, and therefore this second-order Analogy relationship is a marker for a blended space to be created. This blend will be called B ₁ below.
R _{M3}	R ₄ much as R ₆ (I ₅ join[ed...] in an appropriate pattern produces I ₄ and I ₁ /I ₃ join[ed...] in an appropriate pattern produces I ₂ and I ₁)	Analogy	Similarly to R _{M2} above, this compresses to Similarity and creates blended space B ₂ .
R _{M4}	R ₅ [parallels] R ₇ (I ₄ [genitive] I ₁ /I ₂ [genitive] I ₁)	Analogy	As above, creates blended space B ₃ .
R ₁	I ₁ of I ₂ (Since we speak of the 'fabric' of spacetime)	Property	No obvious compression.
R ₂	I ₂ stitched out of I ₃ (spacetime is stitched out of strings)	Part-Whole and Cause-Effect	In B ₁ and B ₂ .
R ₃	I ₄ stitched out of I ₅ (shirt is stitched out of thread)	Part-Whole and Cause-Effect	In B ₁ and B ₂ .
R ₄	I ₅ join[ed...] in an appropriate pattern produces I ₄ and I ₁ (joining numerous threads together in an appropriate pattern produces a shirt's fabric)	Part-Whole and Cause-Effect	In B ₂ and B ₃ .
R ₅	I ₄ [genitive] I ₁ (a shirt's fabric)	Part-Whole	In B ₁ .

Relationship	Input Spaces	Original Vital Relation	Compressed Vital Relation/ Combined Space
R ₆	I ₃ join[ed...] in an appropriate pattern produces I ₂ and I ₁ (joining numerous strings together in an appropriate pattern produces what we commonly call spacetime's fabric)	Part-Whole and Cause-Effect	No obvious compression.
R ₈	I ₆ consists of I ₃ (Matter [...] would then amount to additional agglomerations of vibrating strings)	Part-Whole	No obvious compression.
R ₉	I ₇ composed of I ₃ (Every particle is composed of a tiny filament of energy [...] shaped like a little string)	Part-Whole	No obvious compression.
R ₁₀	I ₈ operates just as I ₃ (And just as a violin string can vibrate in different patterns [...] the filaments of superstring theory can also vibrate in different patterns)	Analogy	Similarly to R _{M2} above, this compresses to Similarity and creates blended space B ₄ .

CIT Table 4: Strings Sample Extract, Potential to Compress

Before a discussion of the blended spaces created by these relations, the relations marked 'No obvious compression' above require some analysis. Leaving R₁ and R₆ aside for the moment, R₈ and R₉ are very similar, indicating that the superstrings of atomic theory are constituents of matter and particles respectively. Given that these latter terms (I₆ and I₇) can be thought of as broadly synonymous, or at the least in a superordinate/hyponym relationship (matter is composed of particles; the *OED2* definition of particle, subsection 2a,

is ‘A very minute portion or quantity of matter’, and an almost identical definition is given in the *Oxford Advanced Learner’s Dictionary* and in *COBUILD*), there is therefore an argument to institute a relationship of Identity between them and so merge them into a single space (I₉). This is the first time in this section that a space merger has been proposed on wholly semantic rather than syntactic grounds, and it is only proposed here near the end of this process, when as much has been achieved as possible through the application of a detailed discovery procedure. With the new I₉, R₈ and R₉ are defunct and must be merged into a new relationship (R_{II}), which states that I₉ consists of I₃ in a Part-Whole relationship.

This leaves R_I, R₆ and R_{II} as relationships which have not yet been identified as having a compressed equivalent in a putative blend. All three are, notably, constitutive relationships, which give extra information about one of their input spaces. In the case of R₆ and R_{II}, both relate to I₃ (*Strings (atomic physics)*), and so can be thought of as adding information to I₃; in which case, both are Part-Whole relationships (I₃ is a part of the whole that constitutes I₆ and I₇) and their compressed equivalents would be, in a resulting blend, Uniqueness (Fauconnier and Turner 2002:97). Note that this reinterpretation of R_{II} and R₆ would necessitate a further second-order relationship of Analogy (compressing to Similarity) to be created between the two relationships due to their parallel definitional structure (R_{M5}). The blend resulting from R_{II} and R₆ acting to add structure from I₆ and I₇ to I₃ is labelled B₅. A similar relationship holds between R_I, adding a property from I₁ to I₂, and in so doing creating B₆.

4.v.4.e.2: First-Order Blend Identification

The first blends arising from these relationships are listed in CIT Table 5.

Blend	Constitutive Relationship	Relationships/Spaces Which Construct Constitutive Relationship	Inner-Space (Blended) Vital Relation
B ₁	R _{M2}	R ₂ R ₃ ; I ₂ I ₃ I ₄ I ₅	Similarity
B ₂	R _{M3}	R ₄ R ₆ ; I ₁ I ₂ I ₃ I ₄ I ₅	Similarity

Blend	Constitutive Relationship	Relationships/Spaces Which Construct Constitutive Relationship	Inner-Space (Blended) Vital Relation
B ₃	R _{M4}	R ₅ R ₇ ; I ₁ I ₂ I ₄	Similarity
B ₄	R ₁₀	I ₃ I ₈	Similarity
B ₅	R _{M5}	R ₆ R ₁₁ ; I ₃ I ₆ I ₇	Similarity
B ₆	R ₁	I ₁ I ₂	Similarity

CIT Table 5: Strings Sample Extract, First-Order Blends

All of the six blends here have Similarity inner-space relations, compressed from Analogy. This is perhaps to be expected with highly analogical texts, or may be an artefact of the procedure of creating these first-stage blends. Regardless, at this stage, it is now possible to begin constructing the first round of blended spaces. B₅ and B₆, being simpler blends which merge properties to create a unique blended space, are the easiest to create (one may note that in a less detailed blending analysis, an analyst would likely have retroactively called B₅ an input space, merging the information given by the referring expressions which created I₃, I₆ and I₇ as properties of I₃). Therefore, for the sake of this exposition, the blends will be approached in descending order (from B₆ to B₁). Their description follows, in the tabular format preferred by Coulson and others. Each column describes either an input space, a relation between spaces, or the blend, while the contents of the input spaces are taken from the description above of the lexical referring expressions of each, combined with the space's label (in italics) taken from the *HT*. The first blend, B₆, is therefore as follows:

I ₁ (Input)	I ₂ (Input)	Inner-Space Vital Relation, R ₁	B ₆ (Blend)
<i>Textile fabric</i>	<i>Space(time)</i>		<i>Space(time) 2</i>
fabric			
	spacetime		spacetime
stitched	stitched	Similarity	stitched

I₁ (Input)	I₂ (Input)	Inner-Space Vital Relation, R_I	B₆ (Blend)
pattern of thread			pattern of [unspecified; empty role slot]

CIT Table 6: Strings Sample Extract, B₆

In a (very) straightforward manner, this blend creates an impression of spacetime as an entity with an analogy to fabric. The key structure is carried from I₂ into the blend, as the structuring relationship, R_I, is of the form I₁ of I₂, giving I₂ syntactic prominence. The entire blend therefore carries information about spacetime and inherits properties from I₁, in particular the phenomenon of patterning. There is no indication of what the spacetime pattern is made of in the blend, and no appropriate relationship to indicate that the literal patterning should be imported in its entirety, however this absent role slot is anticipated by a future second-stage blend, which explicitly addresses this.

The next blended space is B₅, as follows:

I₃ (Input)	I₆ (Input)	I₇ (Input)	Inner-Space Vital Relation, R_{M5}	B₅ (Blend)
<i>Strings</i> (atomic physics)	<i>Matter</i>	<i>Particle</i>		<i>Strings</i> (atomic physics) 2
strings				strings
	matter			
		particle		
stitched				stitched
joined in pattern to form spacetime's fabric		composed of strings	Similarity	joined to form particles joined in pattern to form spacetime's fabric
vibrates can vibrate in different patterns				vibrates can vibrate in different patterns

I₃ (Input)	I₆ (Input)	I₇ (Input)	Inner-Space Vital Relation, R_{M5}	B₅ (Blend)
energy tiny string- like filament of energy				energy tiny string-like filament of energy
	agglomerations of strings		Similarity	forms matter

CIT Table 7: Strings Sample Extract, B₅

Again, this space structures a blend about strings, to clarify further their features. Entities are added to B₅ to create a first-stage blend which gives the reader access to further information about strings.

The next blend is B₄, which is not as simple as B₆ and B₅ were expected to be.

I₃ (Input)	I₈ (Input)	Inner-Space Vital Relation, R₁₀	B₄ (Blend)
<i>Strings (atomic physics)</i>	<i>Violin string (music)</i>		<i>Strings (atomic physics) 3</i>
strings			strings
	violin string		
stitched			stitched
joined in pattern to form spacetime's fabric			joined in pattern to form spacetime's fabric
vibrates can vibrate in different patterns	can vibrate in different patterns different vibrations produce different tones	Similarity	vibrates can vibrate in different patterns different vibrations produce different [unspecified; empty role slot]
energy tiny string-like filament of energy			energy tiny string-like filament of energy

CIT Table 8: Strings Sample Extract, B₄

This blend creates a third variation of information regarding strings. It inherits the violin string property of differing production depending on vibration, but without any explicit replacement for the tone property.

Moving onwards, B_3 is outlined in CIT Table 9:

I_1 (Input)	I_2 (Input)	I_4 (Input)	Inner-Space Vital Relation, R_{M4}	B_3 (Blend)
<i>Textile fabric</i>	<i>Space(time)</i>	<i>Shirt (clothing)</i>		<i>Space(time) 3</i>
fabric				
	spacetime			spacetime
		shirt		
stitched	stitched	stitched	Similarity	stitched
pattern of thread		has fabric		has fabric pattern of [unspecified; empty role slot]

CIT Table 9: Strings Sample Extract, B_3

This appears similar to B_6 , but accounts for a different phenomenon. B_6 , as a result of R_1 , discussed the relationship which held because of the construction ‘ I_1 of I_2 ’, resulting from the clause ‘Since we speak of the ‘fabric’ of spacetime’. B_3 , however, results from R_{M4} ’s second-order relationship of parallelism between R_5 and R_7 , themselves linking ‘ I_4 [genitive] I_1 ’ and ‘ I_2 [genitive] I_1 ’ because of the constructions ‘a shirt’s fabric’ and ‘spacetime’s fabric’. It is as a result of the similarity between R_7 and R_1 that R_7 could be seen here as an unnecessary repetition, and it is only R_5 ’s contribution (the addition of I_4 *Shirt (clothing)*) that differentiates the two. It would likely be preferable for analysis to simply create one blend, linking R_1 , R_5 and R_7 , and so compressing these blends and relationships, however for the present purposes of a detailed worked example, this was not followed.

The next blend is B_2 , which operates across five input spaces, based on the relationship R_{M_3} , which in turn links R_4 and R_6 , relationships joined by the expression ‘much as’. The overall textual stretch this blends is ‘much as joining numerous threads together in an appropriate pattern produces a shirt’s fabric, maybe joining numerous strings together in an appropriate pattern produces what we commonly call spacetime’s fabric’. The relevant profiled section in the blend is the link between the shirt made of joined threads and spacetime’s fabric made of joined strings, and the modalization ‘maybe’ around R_6 means it is the likely organising space. This is outlined in CIT Table 10 below, which requires a full page to show.

The key element here is ‘strings form in pattern to form spacetime’s fabric’. This is a dynamic construction created during the blending process from taking an element from I_3 to unite the ‘pattern of thread’ from I_1 , the ‘has fabric’ from I_4 , and the ‘joined in pattern to form fabric’ from I_5 . This operates in order that, in Fauconnier and Turner’s words, ‘the conceptual complexity of the inputs from several sources is reduced considerably’ (2002:260) to create something which is ‘cognitively manageable’ (*ibid*). The details of the operation are due to the Similarity inner-space vital relation which holds between the various pattern/fabric elements and is derived from the Analogy outer-space relation. This links these five elements and uses a reader’s world knowledge to relate them together. In particular, the three human-scale physical items (shirt, fabric and thread) are easily linked (a shirt is made of stitched fabric, which is composed of a pattern of thread) and only require being related to the structuring space of spacetime to form the blend. Taking the statement above and replacing *fabric* with *spacetime*, and using the Similarity relationship between I_3 and I_5 ’s patterning to determine that strings replace thread in the blend, the conclusion of spacetime being made of stitched fabric, composed of a pattern of strings, is straightforwardly arrived at. (See further §4.iv.5 above.)

I_1 (Input)	I_2 (Input)	I_3 (Input)	I_4 (Input)	I_5 (Input)	Inner-Space Vital Relation, R_{M3}	B_2 (Blend)
<i>Textile fabric</i>	<i>Space(time)</i>	<i>Strings (atomic physics)</i>	<i>Shirt (clothing)</i>	<i>Thread</i>		<i>Space(time) 4</i>
fabric						
	spacetime					spacetime
		strings				
			shirt			
				thread		
stitched	stitched	stitched	stitched	stitched	Similarity	stitched
pattern of thread		joined in pattern to form spacetime's fabric	has fabric	joined in pattern to form fabric	Similarity	has fabric strings join in pattern to form spacetime's fabric fabric consists of pattern of strings
		vibrates can vibrate in different patterns				
		energy tiny string-like filament of energy				

CIT Table 10: Strings Sample Extract, B_2

The final blend is B_1 , which contains many of the same input spaces as B_2 , but focuses instead on R_{M2} , which links R_2 and R_3 (compare B_2 , where R_{M3} relates R_4 and R_6). In so doing, B_1 profiles the verb *stitched* in the previous sentence to B_2 's focus ('spacetime is stitched out of strings... much as... shirt is stitched out of thread'). It is therefore similar to B_2 in that B_2 is a restatement of B_1 (the text extract which B_2 is concerned with immediately follows B_1 's, and begins with the restatement marker 'That is'). B_1 operates as follows:

I_2 (Input)	I_3 (Input)	I_4 (Input)	I_5 (Input)	Inner-Space Vital Relation, R_{M2}	B_1 (Blend)
<i>Space(time)</i>	<i>Strings (atomic physics)</i>	<i>Shirt (clothing)</i>	<i>Thread</i>		<i>Space(time) 5</i>
spacetime	strings	shirt	thread		spacetime
stitched	stitched	stitched	stitched	Similarity	stitched
	joined in pattern to form spacetime's fabric	has fabric	joined in pattern to form fabric	Similarity	has fabric strings join in pattern to form spacetime's fabric
	vibrates can vibrate in different patterns				
	energy tiny string-like filament of energy				

CIT Table 11: Strings Sample Extract, B_1

The outer-space Analogy relationship here maps the I_2/I_3 relationship onto I_4/I_5 . This means that the emphasis here is on understanding the similarity between

the two relationships, so that this blend differs from B₂ by the absence of a ‘fabric consists of a pattern of strings’ element.

There now exist six descriptions of blends, as described in CIT Table 12 below.

B ₁	B ₂	B ₃	B ₄	B ₅	B ₆
<i>Space(time) 5</i>	<i>Space(time) 4</i>	<i>Space(time) 3</i>	<i>Strings (atomic physics) 3</i>	<i>Strings (atomic physics) 2</i>	<i>Space(time) 2</i>
spacetime	spacetime	spacetime	strings	strings	spacetime
stitched	stitched	stitched	stitched	stitched	stitched
has fabric strings join in pattern to form spacetime’s fabric	has fabric strings join in pattern to form spacetime’s fabric fabric consists of pattern of strings	has fabric pattern of [unspecified; empty role slot]	joined in pattern to form spacetime’s fabric	joined to form particles joined in pattern to form spacetime’s fabric	pattern of [unspecified; empty role slot]
			vibrates can vibrate in different patterns different vibrations produce different [unspecified; empty role slot]	vibrates can vibrate in different patterns	
			energy tiny string-like filament of energy	energy tiny string- like filament of energy	
				forms matter	

CIT Table 12: Strings Sample Extract, Blends

A question may now arise as to whether these constitute separate blends within CIT, or if they are instead artefacts of the methodology of this thesis. It is likely that the phenomenon is due to the latter, and that in an impressionistic analysis the various similar blends would be compressed (so that B₁, B₂, B₃ and B₆, which differ very little, and B₄ and B₅, which are similarly compatible, would simply form two blends). As a general principle, then, to remove duplication of blends resulting from the detail of the methodology introduced in this chapter, any blends created which are compatible with another blend can be merged, providing the blends share the same organising input space (such as *I₂ Space(time)* or *I₃ Strings (atomic physics)* in the examples above), have no contradictory elements, share a significant number of common elements, and are at the same level (the blends above are all first-order blends).

This results, for the blends above, in two final spaces, B₇ and B₈. In their construction, B₃ and B₆'s empty role slot is filled by B₂'s role of strings as the constituents of the pattern of spacetime's fabric. The only incomplete element remains the violin/vibration element, which is unanswered in the extract. This point is returned to later. The structure of these blends is as follows (in this table, no similarities are implied through entities sharing rows, for the sake of concision):

B ₇	B ₈
<i>Space(time) 6</i>	<i>Strings (atomic physics) 4</i>
spacetime	strings
stitched	stitched
has fabric strings join in pattern to form spacetime's fabric fabric consists of pattern of strings	joined to form particles joined in pattern to form spacetime's fabric
	vibrates can vibrate in different patterns different vibrations produce different [unspecified; empty role slot]

B₇	B₈
	energy tiny string-like filament of energy
	forms matter

CIT Table 13: Strings Sample Extract, Merged Blends B₇ and B₈

This completes the analysis of the blends which arise from the input spaces and relationships above.

4.v.4.e.3: Second-Order Blends

The analysis could easily end here; B₇ describes the first half of the paragraph above, and B₈ the latter. However, this analysis will instead end by continuing to consider the paragraph as a single unit. In this case, the two first-order blends above need to be combined into a single mental space which contains the full interpretation of the paragraph under analysis. This is relatively straightforward, as the two spaces share similar material, namely ‘strings join in pattern to form spacetime’s fabric; fabric consists of pattern of strings’ from B₇ and ‘joined in pattern to form spacetime’s fabric’ from B₈. These form a straightforward Part-Whole vital relation, while the organising frame will be spacetime, the space most at human scale (this is not immediately obvious when comparing the fabric of existence with subatomic energy filaments, but very small things are particularly difficult to cognitively integrate to human scale, whereas that part of existence which permeates throughout is more immediately perceived on a local scale). The blend therefore compresses as follows:

B₇ (First-Order Blend acting as Input)	B₈ (First-Order Blend acting as Input)	B_x (Final Integrated Blend)
<i>Space(time) 6</i>	<i>Strings (atomic physics) 4</i>	<i>Spacetime composed of strings</i>
spacetime	strings	spacetime
stitched	stitched	stitched
has fabric strings join in pattern to form spacetime's fabric fabric consists of pattern of strings	joined to form particles joined in pattern to form spacetime's fabric	
	vibrates can vibrate in different patterns different vibrations produce different [unspecified; empty role slot]	
	energy tiny string-like filament of energy	
	forms matter	
		has fabric tiny string-like filaments of energy form matter within spacetime such strings join together in a pattern constituent strings can vibrate in different patterns different vibrations produce different [unspecified; empty role slot]

CIT Table 14: Strings Sample Extract, Final Integrated Blend B_x

Under analysis, the entry in B_x's final row, which describes the nature of spacetime as described by the text, is constructed as follows. The organising structure of spacetime contains two statements about strings, namely that 'strings join in pattern to form spacetime's fabric' and that such fabric 'consists of pattern[s] of strings'. These references to strings can then be replaced by

information from the string space, taking along whatever information is required to complete the blend. In so doing, the definition of a string is imported into the blend, requiring only the trivial addition of the fact that matter axiomatically exists within spacetime. This results in the blend.

This then means that the structure of the entire operation is as follows: input spaces I_1 to I_8 , consisting of the separate referring expressions of the entire text, have eleven relationships (R_I - R_{II}) between them, and six metarelations (R_{M_I} - R_{M_6}) holding between those relationships. These metarelations and some relationships lead to six possible blended spaces (B_1 - B_6) which, on analysis, are compressed to two blended spaces (B_7 - B_8) which together can integrate to form a single fully-integrated space B_X , which shares content with all of the input spaces and which arises through the relationships between them.

Diagrammatically, this is:

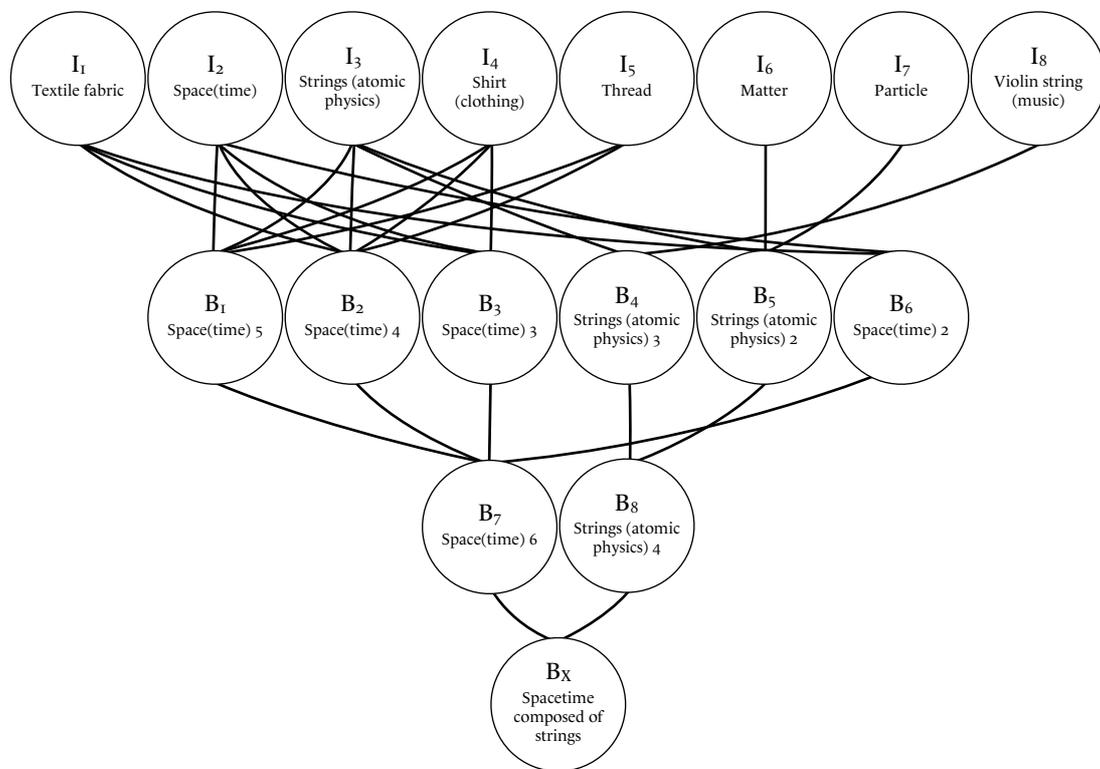


Figure 13: Strings Sample Extract, Blend Relationships Diagram

From this diagram, it is clear how the process has taken a large set of input spaces and relationships and reduced them to a single blend which is composed of a model representation of a reader's understanding of the text. It can be seen from the complexity of the top third of this illustration that the diagrammatic method is not ideal for a CIT analysis of this complexity. Space here does not permit a further layer of relationships between input spaces, as above.

One final point – the empty portion of the role slot describing how strings can vibrate in different patterns but without specifying what this produces (by analogy with violin strings having different tones produced by different vibrations) is not filled here. It is proposed that this is a straightforward strategy to maintain interest; this phenomenon has not been stated previously to my knowledge by any other blending theorist, but it is stylistically highly plausible that an incomplete role slot in a complex blend is a device to encourage the reader to read on further (there is a significant body of work in lexical underspecification in stylistics as an attention-grabbing phenomenon which would support this; see, for example, Leech and Short 1981:219ff, Toolan 2004, Hardy 2005, and Emmott *et al* 2006).

4.v.4.e.4: Summary

Just as the first key domain analysis of *The Fabric of the Cosmos* in Chapter 3 above (§3.ii.1) was heavily detailed and lengthy, the CIT analysis of this single paragraph from the same text has been similarly detailed. While this is a necessary part of describing the methodology of this part of the thesis, it has been of sufficient length to make a summary of the methodological approach advisable:

1. Identify input spaces (§4.v.1). Use referring expressions to identify space builders, use guidance from Relevance Theory to decide which putative space builders are new spaces and which are instead elements to be added to existing spaces. Note in particular the requirement for graded accessibility.

2. Describe the constitution of the input spaces identified (§4.v.2). Use *HT* categories at the fourth hierarchical level of the *HT* ontology to label input spaces, with variations possible for reader clarity.
3. Outline the internal makeup of the input spaces (§4.v.3). Use reasonable schema knowledge to fill out the spaces identified in stage 1 above if necessary, ideally supported by external information.
4. List relationships between the input spaces in the text (§4.v.4.d). Examine prepositional and verbal first-order relationships between the spaces, and then examine metarelations which hold between relationships, by seeing if any prepositions combine processes (eg the underlined prepositional phrase in ‘A stitched out of B much as C is stitched out of D’) or if there are rhetorically parallel constructions which hold between textually-adjacent first-order relationships.
5. Detail the vital relations which each relationship identified in stage 4 above instantiates (§4.v.4.e), first- and second-order. Analyse these vital relations with regards to their compressed equivalents in putative blends. For those relationships with no compressed equivalent, or with no vital relation, consider merging or reinterpreting relationships by adding metarepresentations which rhetorically link the two, only in cases where there is stylistic evidence that a rhetorical link is suggested. (The evidence thus far suggests this may be particularly prevalent amongst constitutive relationships).
6. Construct first-order blended spaces (§4.v.4.e.2). Use CIT guidance to determine the internal structure of a blend as derived from the internal structures of the input spaces (identified in stage 3 above). Combine these first-order blends to form a smaller number of other first-order blends, where appropriate. There will likely be a large amount of repetition in these blends due to the conservative nature of the previous analytic stages. Merge those blends which are ‘compatible’, that is, those

which share the same organising input space, have no contradictory elements, share a significant number of common elements, and are at the same order level.

7. Combine these blends into a final overarching blend (§4.v.4.e.3). Note that this stage may be recursive (eg, five first-order blends might integrate into two second-order blends, which integrate into a third-order overarching blend; there is theoretically no limit to the degree of a final overarching blend).

4.VI: CONCLUSION

This chapter is the longest in the thesis, and is necessarily so for a number of reasons. Firstly, it picks up where Part I left off in identifying narrow sections of text for further analysis, and then it describes in detail the theoretical and methodological approach taken in the remainder of the thesis. The detailed description is necessary, firstly due to the new and changing nature of the field of conceptual integration, and also due to the substantial number of amendments proposed here to make the procedure of blending rather more empirically-based wherever possible, which are best illustrated through a worked example.

The following chapters take this methodology and apply it to the text extracts identified in §4.ii above, to address the final research question of the thesis, regarding reification in popular science.

CHAPTER 5: THE DISCOURSE OF REIFICATION

5.1: INTRODUCTION

This chapter constitutes the core analysis of Part II of the thesis. It takes each analogical extract identified at the start of the previous chapter and proceeds to analyse it in a framework made up of the cognitive linguistic and pragmatic theories outlined above in §4.iv-v. In terms of the thesis' overall argument, this chapter addresses the second hypothesis outlined in Chapter 1, namely that popular science texts present and construe abstractions by means of construing them with reference to multiple and non-congruent concrete domains. Having arrived, via Chapters 2-4, to a stage where it is appropriate to methodologically proceed to test this hypothesis, the discussion below tests the argument with regards to the four texts in Appendix II.2ff, as identified in §4.ii.

The example extract analysed in §4.v.4 above was dense in its construal of an abstract concept, with a large number of domains activated in a short span of (artificially compressed) text. The analyses of the full extracts are not proportional in their density; that is, they are somewhat less dense than one might expect by simply scaling up the size of the sample analysis. This is because the analogical material in these extracts tends to recur and refer back to itself repeatedly, meaning that length of text does not necessarily correlate with the scale of analogical content. Nonetheless, the size of this chapter and scope of the thesis does not permit for any thorough and detailed account of the analytic procedure itself; the exhaustive explicit description of analysis in the example in Chapter 4, in conjunction with the procedural summary of §4.v.4.e.4, will have to suffice for this. Accordingly, the present analyses are mainly displayed in the Appendices related to this chapter (II.2, II.3, II.4 and II.5, covering Kaku's *Parallel Worlds*, Randall's *Warped Passages*, Greene's *The Fabric of the Cosmos* and du Sautoy's *The Music of the Primes*, respectively), with

the exception of a further descriptive sample in §5.ii below. In particular, intermediate stages 4 and 5 of the analysis procedure above are not reproduced in the Appendices, for the sake of brevity. The key elements of analysis – the identification and listing of domains as in stages 1 and 2, a description of their internal makeup as in stage 3, a visual representation of stage 4, and the results of stages 6 and 7 – are all listed in the respective Appendices, and important parts of stage 5 are highlighted in a commentary which follows the analysis.⁴² Overall, this means that the Appendices contain as much information as a reader would wish to have to check how the blending process operates in these analyses, while not sacrificing space for discussion and conclusions in the remainder of this chapter.

Accordingly, this chapter proceeds as follows: firstly, an overview of some of the new methodological points introduced in these analyses, and a short guide to the first analysis as a fully-described example, followed by a commentary on the remainder of the analysis; then, following this, a discursive meta-analysis of the whole example corpus; and finally, some conclusions about the nature of reification in popular science as the analysis here has described it.

5.II: BLENDS AND BLENDING

As described above, this section will offer a commentary on the detailed analysis undertaken of the example corpus identified at the start of Chapter 4. A reminder of abbreviations: *PW*: Michio Kaku's *Parallel Worlds*, *WP*: Lisa Randall's *Warped Passages*, *FC*: Brian Greene's *The Fabric of the Cosmos*, and *MP*: Marcus du Sautoy's *The Music of the Primes*. As there were two methods of identifying analogical content identified in Chapter 4, the following discussion

⁴² Although the analytic description was compressed for concision, it was not desirable to present blended spaces without short explanatory notes, which are appended to the end of each analysis. This is intended to be an effective compromise between concerns of space and explanatory adequacy; a reader who has understood Chapter 4 will have no problem understanding the blending process from the comprehensive list of relevant points in the appendices.

groups together each type of extract, beginning with *PW* and *WP*, those texts which Chapter 3 established had no global key domains but §4.ii.3 identified as having local analogical content as determined by those lexical markers identified in §3.iv. The reason for this is that it is reasonable to expect that the discourse context and construal of temporary and non-systemic analogies may differ from those global analogies in *FC* and *MP*.

5.II.1: BLENDING THE KEY ABSTRACTS

5.ii.1.a: Initial Methodological Notes

This section will briefly run through the process of analysis, beginning with the *PW* text. In §4.ii.3.a above, a 1,017-word extract of *PW* was identified based on analogical markers. This extract is from the closing section of Chapter 7 of the book, titled ‘M-Theory: The Mother of All Strings’, and the foundations of the analogy expressed in this extract are explained in its first paragraph:

Last, there is a rather mysterious prediction of M-theory that is still not understood but may have deep physical and philosophical consequences. This result forces us to ask the question: is the universe a hologram? Is there a ‘shadow universe’ in which our bodies exist in a compressed two-dimensional form? This also raises another, equally disturbing question: is the universe a computer program? Can the universe be placed on a CD, to be played at our leisure?

The two major analogies of the extract follow this, with the first portion of the text discussing the hologram theory, then moving to the computer program theory. These are marked by textual subheadings, dividing the text by analogy. These will be dealt with in turn below.

5.ii.1.a.1: Typographical Conventions and Blend Pre-enactment

The text is reproduced in Appendix II.2.a, annotated with the convention introduced above that **boldface** indicates a new mental space, *italics* an

anaphoric⁴³ referring expression to a prior space, and underlining a lexical, phrasal or clausal element to be integrated into a previous mental space, or a lexical element which constitutes a nominal premodifier. This means that the rare uses of italic type for emphasis in the texts in this chapter are removed in the annotated versions, although are retained in the full extract which starts each appendix. On those occasions elsewhere in this chapter where there are two mental spaces created adjacent to each other, a vertical bar (|) is used to separate them, to distinguish between different mental spaces activated by adjacent NPs and those single spaces activated by an noun-noun compound.

One further typographical marker is required. The reason for this is that there are two instances of what will be called *blend pre-enactment* in this extract; instances where a resulting in-text blend is cataphorically referred to ('By *analogy*, X', where X is the remainder of the paragraph outlining an analogy, and the word *analogy* therefore refers to the outcome of the blend which is being set up; similarly, 'there is a rather mysterious prediction' is a predefining label for the blend which is being set up in the passage). These are not uncommon devices in text, although are rarer than anaphoric references (see Halliday and Hasan 1976:17ff for a lexicogrammatical perspective on this), and are therefore occasional exceptions to the general outline above that the first indicator of a mental space is highlighted as a space enactor rather than a simple reference to a space (as a blend is only constructed after input spaces are identified and blended, it is not possible to describe these cohesive devices as enactors of an as-yet non-existent space; thus the term pre-enactment introduced above). These cataphoric indicators are therefore both italicized and suffixed by a greater than symbol, >, to aid a reader (anaphora is simply marked by *italics*). As an aside, it may be hypothesized that such discourse cataphoric

⁴³ *Anaphora* and *cataphora* are used in this chapter and elsewhere solely to mean direction of reference to the same mental space. Halliday and Hasan 1976 would strictly term this phenomenon *lexical cohesion*, as anaphora in their system requires the use of particular types of lexical expression. Therefore, while a reference is made to Halliday and Hasan's theory of cohesion below, this thesis is focused on mental space anaphora/cataphora, irrespective of enactor type.

reference devices are particularly associated with blends in complex discourse due to the principle of end-weight in English (Leech and Svartvik 2002:158), given that blends require longer and more syntactically-complex discourse than does a reference device.

5.ii.1.a.2: Reduction

This section outlines some reductions to the texts in the appendices which have been carried out for the sake of concision.

Firstly, parenthetical statements (full sentences enclosed wholly in brackets) are ignored as axiomatically discoursally parenthetical. However, sentence-intermediate clausal or phrasal elements in parentheses are retained in the text as possible meaning-relevant supplementary interjections, constituting supplementary elements in an existing mental space, as opposed to full parentheses. Relevant here is the condition that listing parentheses (eg ‘such as the information problem, calculating the masses of the quark model, and so forth’) likely constitute elements in an example mental space, rather than a series of mental spaces in their own right.

Also omitted, for the sake of brevity, are mentions of scientific attributive agency. This includes factive and non-factive metarepresentations (for a description of metarepresentations in discourse see Zunshine 2006, Chapter 3) such as ‘Some cosmologists have conjectured’ and ‘Bekenstein and Hawking conjecture’, in addition to attributional adverbial clauses such as ‘In 1997, Juan Maldacena, at the Institute for Advanced Study at Princeton, created quite a sensation when he showed that’, where the main verb in the attributional clause is a factive which links circumstantial information to the main clause. The reason these can be omitted in this case is that there are no contrastive elements in this text extract where one set of metarepresentations or circumstantial attributives is set against contradictory or opposing elements. (So it would not be appropriate to remove these metarepresentations, for example, where a text states ‘Newton theorized [...] but later, Einstein proved

this to be untrue’, where the juxtaposition of contrast changes the discourse function of these elements from circumstantial to substantive.) In cases where such features are textually circumstantial, the identified mental spaces and elements within agency statements are removed from the annotated texts.⁴⁴ This rule is applied somewhat conservatively, permitting existential clauses with dummy subjects with arguably elided circumstantial agency, such as the existential particle in ‘Last, there is a rather mysterious prediction of M-theory that is still not understood’ (so the nominalization *prediction* is retained and therefore constitutes an initial mental space, whereas, by contrast, ‘X and Y have predicted’, would be omitted; arguably such nominalizations could also be ignored, just as the implicative or factive verb would be, but this would result in a possible overreach of a simple rule designed to condense textual exposition). This does result in occasional anaphorically-orphaned pronominals, which are retained for textual coherence but are not treated as possible mental space enactors. Similarly, as in Chapter 4, rhetorical uses of generic pronouns, such as ‘we’ as an inclusive audience marker, or metatextual noun phrases (‘in this book, we will’, etc), are ignored with respect to mental space formation. Finally, there is some small flexibility over questions of grammatical gradience, for example in those instances where there are numerals which are arguably either determiners in NPs with elided heads, or being used substantively as NPs in their own right. These are dealt with on a case-by-case basis. All these analytic rules serve to reduce the number of extraneous (that is, not to do with the main informational focus of the text) mental spaces under analysis; while it would be possible, and actually even simpler, to account for the status of every referring expression in text, the space requirements of this thesis and the limits of this chapter in particular mean that it is preferable to reduce the analytic spread as much as possible, for the sake of

⁴⁴ This is not to argue that metarepresentations and other circumstantial information do not figure substantially in a mental representation of a text, but rather that they operate as a distinct layer of representation which overlays but does not, in this non-contrastive use, interact with the main informational content of the discourse (Zunshine 2006).

exposition. In many cases, it is arguably more straightforward to account for every such expression, should there be unlimited space for analysis, than to apply these reduction principles. Readers are, however, referred to the somewhat protracted analysis in §4.v above as a sign of the expository volume which would be necessary for an exhaustive examination of these texts (in length, approximately 80 times greater than the example in §4.v), which is not possible or desirable in a thesis of this nature.

5.ii.1.b: Analysis

The Appendix for *PW* proceeds as follows: after the full extract is quoted in part (a), the reduced and annotated version is provided in (b). Following this are tables and lists, concluded by the notes and commentary on the blending process. The mental space table in part (c) of Appendix II.2 therefore lists all the input spaces for this extract. Notable here is the *HT*'s omission of *M-theory*, which was omitted from *OED2* and the *Additions* series. (This is understandable, given its very recent coinage; it is, however, in *OED3*.) I have here assigned it to the existing category 01040707: *Quantum theory*. The *anti-de Sitter universe* and *de Sitter universe* enactors were similarly assigned to 011010: *Universe*. Wherever this occurs in the remainder of the analysis (which is rare), I have used my own judgement as a former section editor of the *HT* to decide where best a neologism or specialized term should be placed. None of these decisions, to the best of my knowledge, would be considered controversial by another *HT* editor.

In addition, where a space already exists and another form of the same idea is instantiated but with a modifier – such as *universe* and *shadow universe* – then the textual premodifier has been added to the name of the new input space, to help distinguish it. As there are a number of multi-lexemic enactors in this space, further manual searches of the *HT* were necessary in some cases. Blended spaces are given names either based on their main structuring space or just named after a straightforward textual cue which suggests what the blend should be considered as (prefaced by X rather than an *HT* category number). In some

cases, cataphoric reference (as described in §5.ii.1.a.1 above) results in a blend name being pre-specified, although as this is not an input space it is not given an *HT* category number and is therefore prefixed by X).

5.ii.1.b.1: Groupings

The volume of input spaces is such that it is preferable to attempt to construct some initial blends separately, where possible. There are two grounds on which this is carried out; semantic and textual.

5.ii.1.b.1.a: Semantic Groups

Part (d) of Appendix II.2 lists proposed semantic groups of these input spaces. These are derived as a result of using the *HT* categories for space naming (omitted in the short example in §4.v above, but necessary in the current analysis). This is because, by taking the *HT* category number for related concepts and moving upwards in the hierarchy, where appropriate, semantic categories can be placed in hyperordinate categories of which they are already a part. Certain categories, of course, will not have any semantic neighbours, and will exist in a grouping of their own.

The groups for this *PW* extract, with group names in CAPITALS, are listed in part (d) of the appendix (also note the lack of a colon in a group name, to represent the fact that the connection here between the *HT* category number and the group name is a very loose one). These groups function as representations of the text at a slightly higher level of semantic granularity than the short extract analysed in Chapter 4, for the sake of assisting analysis (see further §4.iv.8.c above for a discussion of granularity).

5.ii.1.b.1.b: Textual Groups

Following from this, there are three main textual indicators of possible space groupings:

1. Text organization. A single paragraph or otherwise functionally-separable discourse segment with a series of mental spaces only activated in that segment will indicate that there is a process of meaning construction within that paragraph which is self-contained. This does not mean that, for example, the paragraph is irrelevant to the surrounding discourse, but rather indicates that the process of meaning construction within that paragraph has internal relevance only, producing a single blend which would then interact with the remainder of the text. Therefore, in the *PW* extract, there is a paragraph of analogy with fish and fish bowls which is connected to the remainder of the discourse by invocation of the explicit term *analogy* and by mentions of holographs and dimensionality. This paragraph can be usefully treated apart from the main text as a rhetorical repetition for expository purposes.
2. Analogical markers. Just as the functionally-separable paragraph of exposition regarding fish and fish bowls above was introduced by explicit reference to the stylistic analogy it contains, other discourse segments which are introduced by analogical markers can be treated as separable groups of mental spaces. In the *PW* extract, there are sentences which evoke series of mental spaces which are introduced by markers such as the conjunction ‘as if’, which function as an indication that the following material is analogical, and can therefore be treated as a separable group.
3. Dense enactor clusters. A series of enactors clustered in a single sentence, with syntactic relations between each, indicates close relations between the enactors which could lead to them being usefully grouped together for analytic purposes.

These indicators are not exclusive; nor are they found in isolation. In contrast, there are in the texts under analysis in this thesis separable discourse segments

introduced by an analogical marker and consisting of dense clusters of mental space enactors. Such co-location of grouping indicators reinforces the case for them to be considered appropriate groups for separable analysis. For the present text extract, there are three such groups: the ‘fish’ paragraph referred to above, the window/keyhole ‘as if’ sentence, and the dense co-ordinated list ‘proton and neutron’.

5.ii.i.b.i.c: Grouping and Compression

Both the extract’s semantic groups and the textual groups also identified are therefore listed in part (e) of Appendix II.2; later appendices omit the full list of groups and instead include them directly in the table.

Based on these groupings, certain spaces can be compressed into a single space. Firstly, the clause introduced by the analogical marker conjunction ‘as if’ contains two interchangeable examples which can be schematized to a higher level (‘as if you were looking at the *image* through a **window** or a **keyhole**’).

This is supported by the evidence that they are both within the same semantic category and a textual cluster, as identified above. We can therefore move one level up in the *HT* hierarchy and name this a single space 01050705 *Aperture* (after the precise shared category 01.05.07.05.04.01 *An opening/aperture*⁴⁵).

Similarly, in the clause ‘One goal is to calculate the masses and properties of the **proton** and **neutron**’, *proton* and *neutron* can be collapsed into 01040705 *Baryon*, its hyperordinate term (cf *HT* category 01.04.07.05.04.06.02 *Baryon*, within which the subcategories of *proton* and *neutron* sit). Part (e) of Appendix II.2 therefore shows the resulting list of 24 input spaces, with groupings noted. Later appendices do not describe this step, but apply it directly to the Grouped Mental Spaces table.

Finally, note that, unlike in Chapter 4, the *HT* labels for each space are retained in their names in these analyses for reasons of aiding this grouping. This makes

⁴⁵ Note that space names, as explained in §4.v.4.c above, have no intermediate periods in their L4 category reference number, whereas references to actual *HT* categories do.

the names longer, however, and so in the relationships diagram (see §5.ii.1.b.c below) they are omitted for readability reasons.

5.ii.1.b.2: Grouped Blends and Elements

Following this compression, the grouped blends are associated with their main elements in Appendix II.2 part (f), and these elements have been paraphrased where necessary to aid reader comprehension. Certain input spaces have no elements defined by the text, but rather operate solely as instantiations of either schema knowledge or prior discourse knowledge/blends - what Sanford and Garrod (1981:159; 1998:162) would call 'Memory for the Discourse'. While all mental spaces are supplied with schema knowledge, it is comparatively rare in this analysis for spaces to be used with only schema knowledge, and no textual elements. This is perhaps because of the nature of those areas of these texts under analysis, which feature a particular breadth of focus, rather than depth, and the selection procedure in Chapter 4, which used textually-signalled topic shifts for extract identification. A further reason is that schema knowledge may not be as necessary in expository text as it is in narratives, as this discourse type⁴⁶ is often aligned towards information rather than narrative progress.

Note that, in the Mental Spaces and Elements part of each appendix, there will be elements in the early mental spaces which are from later parts of the discourse, and often, due to the nature of topic-focus, elements which are part of the extract's conclusion will be added to the initial mental spaces (for example, *PW I₁*). Finally, necessary information added to a mental space from putative schema knowledge, as outlined in §4.v.3, is placed in *italics* in the analytic appendices (eg *MP I₂*).

⁴⁶ For a discussion of the concept of 'discourse type' in linguistic and stylistic work, see, *inter alia*, Carlota Smith (2003) and Egon Werlich (1976). The distinction here is based on the definitional criteria in Smith (2003:12ff).

5.ii.1.b.3: Relationships

Again, given limits on space, there is no possibility of listing all relationships and metarelations between all the input spaces in the extracts. Instead, a simple visual guide to relationships between spaces is provided in (g) of the appendix, which also usefully functions as a summary of the input spaces and their progression in the text. These diagrams will be returned to later in this chapter for the purposes of analysis.

5.ii.1.b.4: Blends and Commentary

Each analytic appendix ends with the resulting blends for each extract, listed in a table with their source domains and the blended elements. Those spaces with important emergent structure have such emergent elements highlighted in **bold**. Each extract has one or more first-order blends resulting from the inputs, and all have a further B_x blend which is either a second-order or third-order blended space. The B_x space represents the overarching mental space which the text extract has as its ostensive expository goal; their status as the ‘point’ of a text is a slippery and problematic one, although is used in the blending commentaries as a form of shorthand for the nature of an overarching blended space (or ‘arch-blend’). A commentary on the blending process is also provided, for clarification of certain critical points.

5.II.2: DETAILED NOTES

The description above focused on *PW*, as a detailed example. Below are some short notes about the other analyses, to be read in conjunction with the Appendices.

5.ii.2.a: *Warped Passages*

This begins with a short description of a ‘Wonkavator’ (from Roald Dahl’s *Charlie and the Chocolate Factory*), unsurprisingly not present in the OED, which I have assigned to its most pertinent *HT* category (03.09.02.02 *Conveyance*). The

same has been done to *hypersphere* and *hypercube*, both assigned to 01.05.07.03 *Shape*. The ‘a fictitious two-dimensional universe’ enactor, with its premodifiers, necessitates a narrowly-defined space, called here 010110: *Flatland Universe*. Conversely, ‘our world’, being unmarked for deviance from the norm, is simply categorized as 010110: *Universe*.

The character name ‘A. Square’, from Abbott’s 1884 *Flatland*, is particularly interesting, as it is simultaneously both a proper name (and therefore not represented by an *HT* category) and a description of the being (literally, a square). I have given it its pertinent *HT* category for the referent of the name (01.05.07.03 *Shape*) but prefixed it with an X to indicate it is simultaneously a proper noun. This name is, of course, also a further blend of function and identifier, but such onomastic description is outwith the bounds of this thesis. Similarly, the unnamed five-dimensional ‘something’ in the penultimate paragraph of the extract is given the same category (01.05.07.03 *Shape*) but with the prefix to indicate that this assignment is somewhat questionable (lexicologically speaking, there are philosophical issues surrounding the addition of a non-lexemic entity to a hierarchy based on lexemes).

Borrowing a convention of Text World Theory and Possible World Theory (see Werth 1999, Gavins 2007 and Ryan 1991), those mental spaces which are part of the description of the alternate universe of Flatland are considered here as embedded in an alternate instance of factuality from the perspective of the main discourse (one might add that this could be seen as a new text world, but it is not relevant to this thesis to integrate Text World Theory into the present exposition; from some perspectives, the theories could be seen as broadly congruent at different levels of granularity). This is textually signalled by the perspective shift devices ‘a fictitious universe’ and ‘our universe’. A further reference is made to the fictional universe later in the text (‘it existed in Abbott’s story’, where *story* is taken here as an anaphoric reference to the Flatland universe).

5.ii.2.b: *The Fabric of the Cosmos*

This relatively short extract requires no further notes other than those given above and in Appendix II.4.

5.ii.2.c: *The Music of the Primes*

The space which refers to the axis marking the line where the graph of the zeta function is 1 has no *HT* category, and so is referred to as the ‘1-line’ in the analysis, with the *HT* category 01.05.07.05 *Relative position*, prefixed, as with *WP* above, with an X to indicate its marginal category status.

5.II.3: MULTIMODALITY

Finally in this section, a note on what systemic-functional analysts call ‘multimodality’ (the use of multiple modes of communication, generally including both the visual and the textual). This is used in three of the four texts, with diagrams supplementing the text in occasional places. All such diagrams are reproduced faithfully in the appendices. The following discussion outlines why treating these images as supplementary elements in a CIT analysis is an acceptable way to handle them for the purposes of this thesis. This is in contrast to many theorists in other fields of linguistics, such as van Leeuwen and Jewitt (2001), or Kress and van Leeuwen (2001, 2006), for whom visual semiotics are treated as necessarily distinct from the semiotics of written text, and are customarily analysed as fundamentally different from their accompanying text (so, for example, systemic-functional analysts employ the same approach towards visual images as a social semiotic as they would to linguistic items, but do not use identical analyses and instead tailor their analytic toolkit towards images). This may well be a required viewpoint within systemic-functionalism, whose analyses are necessarily language- and mode-specific, but is not necessarily the case within cognitive semantics.

Theoretically speaking, the cognitive approach to both semantics and grammar, as characterized by the influential theorists of what we may call the ‘California

school',⁴⁷ has generally converged around two core 'commitments' which dictate and motivate research within the paradigm (Lakoff, 1990:40-47), both of which are useful to determine, in theory, the cognitive approach to multimodality. Cognitivists do not discuss 'multimodality' using that term, if at all – the theoretical considerations of the movement are primarily linguistic almost to the exclusion of visual considerations. Of course, the cognitive linguistics enterprise is relatively new (its birth as a cohesive theory can be traced to the 1989 Duisburg conference and the 1990 founding of the journal *Cognitive Linguistics*⁴⁸) and the absence of analyses involving the visual channel of communication may be taken not as theoretically precluded but rather as being in an area not yet explored by analysts. The following discussion will attempt to explain why there are no theoretical barriers to a 'multimodal' discussion of texts within cognitive linguistics.

5.ii.3.a: The Lakovian Principles

According to the first of Lakoff's two core principles, cognitivists are committed to *generalization* – the movement towards general principles which are applicable to all of language (1990:40). This is why cognitive theories of semantics, syntax, discourse and morphology employ the same models at differing levels of granularity, and why conceptual integration can be used simultaneously at both a general cultural/discoursal level and at the morphological (and, rarely, at the morphophonemic) level. A significant part of the history of the movement, in the pre-Duisburg days, was as a counter-reaction to the Chomskyan approach to a modular theory of both linguistic analysis and language production (Kemmer 2010 and Geeraerts and Cuyckens 2007a:4ff). Thus the commitment to identifying a common set of cognitive abilities inherent in mental processing which the human language facility

⁴⁷ George Lakoff, Ron Langacker, Seana Coulson, Charles Fillmore and Gilles Fauconnier are based at the various campuses of the University of California, although Mark Turner is based in Ohio, Len Talmy in New York, Adele Goldberg at Princeton and Mark Johnson in Oregon.

⁴⁸ For more details on the history of the discipline, see Langacker 2002: ix-xiii and Evans, Bergen, & Zinken 2007.

draws upon in its operation can produce a common set of concrete analytical methods which are applicable to a wider range of phenomena than would be expected within the formal or functionalist tradition. This is applicable to the current discussion in a number of ways, but the two most significant applications are the linked fundamental dedication to a breadth of analytic objects and the refusal to modularize processing tasks. There is here no theoretical barrier to multimodal studies, or indeed to the application of cognitivist theories to a wider semiotic analysis.

The second commitment of Lakoff's paper is the *cognitive commitment* (1990:41), wherein the linguistic principles which are derived from analysis (or applied in the process of analysis) must be those which correspond with general cognitive principles. In this way, cognitive linguistics cannot postulate theoretical structures or requirements which contradict 'what is known about the mind and brain from other disciplines' (Evans, Bergen, & Zinken, 2007:4). To therefore separate the analysis of the construal of multimodal text from the purely verbal or the purely visual would require the multimodal text to be cognitively separate from its parts, which has as yet no evidential requirement (Stenning, 2002:53). Given the current state of knowledge about the mind and brain, it is not reasonable to posit limitations which are not psychologically required – as Dabrowska says (2004:205), the whole point of linguistic forms is to express meaning. Cognitivists take a rather broad view of meaning as expressing not just 'content' but also 'imagery', although this is muddled somewhat by the tendency of cognitive linguists to use *image* to mean both mental image and visual image.

5.ii.3.b: Prior Usage in Conceptual Integration

Moving to more concrete evidence, Fauconnier and Turner (2002:65-67) discuss an image-based advertisement, in which children are shown about to conduct a heart bypass (the intent is to emphasize the importance of education to the future of a country) without any reference to the fact that one part of the

blend is visual and others are verbal. This is likewise reflected by Coulson's article (in press) on political cartoons, which analyses cartoons on what appears to be a purely semantic and semiotic level – she does not specify that some of her input spaces in each cartoon are visual (images of US presidents George Washington and Bill Clinton, and in the latter case an image with lipstick marks on Clinton's face, uncaptioned) and some are textual (titles, slogans or speech bubbles). Todd Oakley (1998) similarly undertakes a lengthy analysis of a graphic novel, Art Spiegelman's *Maus*. It is clear that Fauconnier, Turner, Coulson and Oakley (*inter alia*) do not recognize a distinction between the different modes in play, either within mental space approaches or Conceptual Integration Theory – either because the distinction is trivial in the cognitivist approach or because it is nonexistent, as is argued here.

5.ii.3.c: Conclusion

In summary, it is clear that the cognitivist approach does not preclude any sort of visual or diagrammatic analysis. Indeed, as Langacker (2002:12) says, 'lexicon and grammar form a continuum of symbolic elements'. This continuum can, for our purposes, be extended in further directions beyond the lexicon towards more abstract symbolic or semiotic elements, forming a cline from words and pictograms to diagrams and images. In many ways, the use of multimodality can be seen as an aid to running the blend – providing visual support for the process of going through a mental simulation which the process of blending necessitates. Regardless, it is clear that there are no theoretical or practical justifications for treating diagrams as special or requiring any detailed further analytic apparatus to deal with in the analyses in this chapter.

5.III: REIFICATION IN THE KEY EXTRACTS

5.III.1: OUTLINE

The analyses in Appendices II.2-5 give an insight into the process of reification in popular science, and as samples representative of analogical content in abstract science (see Chapter 4), they constitute an appropriate corpus of reificatory blends in this genre. The following observations and results are not intended to be exhaustive; the nature of the discourse and the range of what constitutes popular science as an overall genre (cf §I.iii) make this a difficult task. Nonetheless, the discussion below is valid across the spread of the present subgenre as a representative exposition.

Three sets of conclusions and comments can be drawn from a consideration of the analytic results as a whole, which are detailed in turn below. The first explicitly addresses the nature of concretization, while the following two offer some commentary on implications of the data more generally. For the ease of cross-comparison, each mental space is prefixed below with the two-letter identifier of each text, giving, eg, WPB_x, FCI₃, MPB₄, etc.

5.III.2: REIFICATION

Reification, as outlined in Chapter 1, is the process of making an abstraction somewhat more concrete. Accordingly, an analysis of blends based on their relative ‘concreteness’ can add to this discussion substantially. The following discussion gives data on this for all texts involved, before an analysis of what this data tells us.

Of *PW*'s six blends, four are straightforwardly concerned with abstract concepts (duality, reality, the information problem, and mathematical problems more generally), while the other two are entities which are extremely far from human scale (the universe and its dimensionality, and a black hole).

Conversely, of its 23 input spaces, all but five are from *HT* section 01, *The*

External World, and overall there are a large number of spaces which are concrete objects at human scale – photographs, objects generally, books, goldfish bowls, fish, holograms (and note that the hologram space explicitly includes an element specifying where they can be found at human scale; credit cards, museums and amusement parks), computer programs (interacted with at human scale), and so on. The figures are that 13 of 23 inputs are abstract, while 6 of 6 blends are, giving an approximate ratio of 1:2 and 1:1 respectively. For *WP*, 6 of 15 inputs are abstract while both blends are abstract, for a ratio of 2:5 and 1:1. For *FC*, 6 of 12 inputs are abstract while again all three blends are, for ratios of 1:2 and 1:1. Finally, for *MP*, 13 of 27 inputs are abstract (for another 1:2 ratio of abstractions to total) while the blends are difficult to categorize. All five of them have elements at human scale (treasure, hills, troughs, and general landscape elements) while also incorporating abstract elements and import. This can be contrasted to PWB_x , WPB_x , and FCB_x , all of which have elements which are either abstract or not at human scale.

Two situations are thus occurring here. Firstly, in three texts, approximately half of inputs are concrete while no blends are, giving a straightforward result of concrete input domains being mixed with abstract domains to form abstract conclusions, as hypothesized. Secondly, in the *MP* text, there is a disanalogous situation with blends being of mixed type, having both concrete and abstract elements. An explanation of this takes into account the explanatory nature of the genre, and of the textual status of both extracts. *PW*, *WP* and *FC* are clearly textually coherent as single units, whereas *MP* breaks off in the middle of a longer analogy. A reading of the remainder of this chapter reveals the final textual conclusion is entirely abstract. While not wishing to dismiss too quickly the possibility of *MP* as instantiating a different style of exposition, any reading of the text must take into account its incompleteness (perhaps leading to the hypothesis that a final wholly-abstract blend is an expected requirement of popularized abstractions), and so categorising MPB_x as actually a further input space in a larger arch-blend is plausible. This leads us to the general conclusion

that, based on this small sample, input spaces tend to be roughly equivalent in their abstract/concrete balance, whereas blends which explicate abstractions are fully abstract, but arising from concrete conceptualizations. The genre-specific hypothesis of this thesis, that popular science texts present and construe abstractions by means of construing them with reference to multiple and non-congruent concrete domains, has therefore been demonstrated, by means of the prevalence of concrete domains as input and the entirely abstract nature of the blends, as well as the textual and semantic spread of concrete domains as outlined in the analytic appendices.

One final note here is that future work on this style of analysis will be greatly complemented by the completion of the forthcoming University of Glasgow *Mapping Metaphor* project (2012-2015), which aims to empirically investigate the foundations of metaphor in English. As a result, information on the nature of concrete/abstract domains will be produced, which will allow for automatic analyses of the proportions of concrete/abstract mental spaces in a text.

5.III.3: TYPES OF BLENDED SPACE

The following table schematizes a typology of those blends created in the process of analysis of the four extracts.

Blend Identifier	Space Name	Type
PWB ₁	01011006: <i>Black Hole 2</i>	Redefinition
PWB ₂	X: <i>Duality</i>	Original, pre-enacted
PWB ₃	X: <i>Analogy</i> [=01050102: <i>Reality 2</i>]	Redefinition, pre-enacted
PWB ₄	01060402: <i>Information problem 2</i>	Redefinition
PWB ₅	01060402: <i>Mathematical problems 2</i>	Redefinition
PWB _x	X: <i>Prediction</i> [= 010110: <i>Universe 2</i>]	Redefinition, pre-enacted
WPB ₁	X: <i>Higher-Dimensional Movement as Change in Intersection</i>	Original

Blend Identifier	Space Name	Type
WPB _x	X: <i>Analogy [Dimensional Passage]</i>	Original, pre-enacted
FCB ₁	OIO50702: <i>Dimension 2</i>	Redefinition
FCB ₂	OIO102: <i>Geodetic reference 2</i>	Redefinition
FCB _x	OIO110: <i>Universe 2</i>	Redefinition
MPB ₁	OIO60407: <i>Zeta function 2</i>	Redefinition
MPB ₂	OIO40810: <i>Three-dimensional shadow 2</i>	Redefinition
MPB ₃	OIO10404: <i>Landscape 2</i>	Redefinition
MPB ₄	OIO10901: <i>Map 2</i>	Redefinition
MPB _x	X: <i>Reimann's landscape</i>	Original

CIT Table 15: Blends in the Key Extract Corpus

If a blend inherits substantial structuring space from an input, then it has in the analyses been identified as a redefinition of that input; essentially, an updated mental space which blends that input space with other input spaces introduced in later parts of the text. However, not all blends are like this, and some (eg MPB_x and the *WP* blends) are new structures, indicated as original.⁴⁹ These arise in two situations in the key extract corpus. Firstly, they characterize situations with significantly equivalent analogical inputs, which compress to Identity in the blend. This occurs in PWB₂, which instantiates the concept of duality as an identity between two dimensionally-defined universes, and also in WPB₁, which again has a vital relation of Identity between two spaces, with extra information from a further input. Secondly, they characterize arch-blends, the final outputs of texts, which have to deal with situations with relatively large numbers of multiplex inputs: these are WPB_x and MPB_x, which

⁴⁹ The term 'Original' is not ideal, as blends are by definition not original but rather arise from existing inputs. There are no terms which are ideal to capture the fact that a blend is not a redefinition but still arises from other inputs; this is due to the slightly paradoxical nature of referring to a blend as a separate creation. 'Original' will have to suffice, with the above caveats.

have five and eleven inputs respectively when prior input spaces to input blends are taken into account. This result, an identification of two particular ‘types’ of blended space, is only possible because of the detailed naming and structural conventions for input spaces outlined in Chapter 4, in particular the use of *HT* data to define the scope of a space. No claims can be made as to the comprehensiveness of this within a further ontology of blends, but this three-part schema – redefinition blend, original blend from equal-weighting of inputs, and original blend from multiplexity of inputs – is a useful result.

In particular, it is useful because it gives an insight into the textual style of each extract. *WP* as a text is particularly relevant here, as both of its blends are original, and this can be attributed to its significant use of Identity as an expository strategy; the full text has two major textual groupings which represent separate worlds (Flatland – a two-dimensional existence, physically speaking – and that of the hypersphere and hypercube – a four-dimensional existence) which parallel each other in order to exposit dimensionality shifts from the three-dimensional real world. This parallel can be seen in the relationship diagram for *WP* (Appendix II.3.e), where, uniquely in the key extract corpus, a blend is required to connect the two textual groupings into a single blending mechanism (see further below on this point). This use of Identity is reflected in the blend types which *WP* employs, free from prior redefinition of individual spaces due to its two-sided equal-weight strategy.

A final point on textual style is the complexity of the MPB_x blend, which, given its expansive set of inputs, could be suggested as an archetypal arch-blend. Notably, all the extracts here analysed are relatively brief when compared to a full book-length exposition, and so it could be expected that arch-blends which represent whole chapters, or whole books, could be yet more complex. In such cases, the emergent structure discovered (and the previously-emergent structure imported from intermediate *n*-order blends) would be so dominant as to make complex arch-blends typically original, as defined here. It can be

hypothesized, and borne out by the limited data available, that such original blends characterize complex and recursive blending mechanisms, as these would be naturally quite divorced from their original inputs. Abstract popular science texts would thus axiomatically require original blends as an output, rather than redefinitions, corresponding with the discourse aim of introducing radically new concepts to a reader (the purpose of popularization), rather than update existing concepts; compare this with marginal popular science texts such as Bryson's 2010 *At Home*, which aims to deepen a reader's knowledge of the history and science of an ordinary home, where the text's main function is to take existing concepts and conceptually develop and refine them.

Secondly in this section, there are some spaces which are pre-enacted (see further §5.ii.1.a.1 above), and textually-signalled with empty space markers. These four are represented by the lexical items *duality*, *analogy* (x2) and *prediction*. While again there are caveats to be borne in mind from the size of the present corpus, it is a reasonable hypothesis that pre-enactors are established by terms with relatively wide reference and at a high level of abstraction. This is because such placeholder markers require flexibility in later interpretation while still being precise enough to actually enact spaces. These could even be considered, in future work, to be signals of generic space types which a reader would treat as priming signals for possible integration networks. In the present corpus, two enactors come from the abstract 01.06 *Relative properties* section of the *HT*, while the third is from 02.01.14 *Expectation* in the mental world section. Future work can expand this list, noting which concepts pre-enactors cluster around, and what their semantic prosody is.⁵⁰

⁵⁰ Four of the top three shared collocates of these terms in the BNC are the abstract lexemes *argument*, *model*, and *theory*, suggesting their semantic prosody is also abstract. (*Argument* here is assumed to be in the sense of a line of reasoning, rather than a verbal altercation.) Note further that the term *semantic prosody* is a variable one in corpus linguistics; it is here used in the same sense as Sinclair (eg 1991:70) and Louw (1993), but without the particular lexicological function for which they employ the technique.

5.III.4: TEXTUAL PROGRESSION MARKERS

With the exception of the *PW* case discussed in §5.iii.2 above, no mental space is isolated in the diagrams in the analytic appendices. This may be an expected fact due to the nature of blending, but it reveals an important possibility for analysis; the use of what I here call *progression traces*. This terminology is adapted from the cognitive linguist Mary Ellen Ryder (2003), who built on the contextual frame theory of Catherine Emmott (1997, 2003) to develop a theory of *character traces*, diagrammatic lines which represent a character's context throughout time (Ryder 2003:213-4). In a similar way but at a higher level of abstraction, a progression trace monitors diagrammatically a text's movement throughout the text's expository progression, as here.⁵¹ The overall effect is represented in the relationship diagrams in each analytic appendix, which are ordered temporally from left to right where possible. Treating the general sense of left-to-right movement as a sign of expository progression, the diagrams fall into two separate categories; roughly linear, as in *WP* and *FC*, and rather more matrical, as in *PW* and *MP*.

For the second category, these matrix-like diagrams can further be described in one of two ways. The first of these primarily describes excursus, as in *PW*, where the central textual matrix does not progress from one textual grouping to another as in the linear examples above, but rather 'dips' in and out of three textual groupings, mirroring the text structure, which moves through three explanatory domains (the black hole, the anti-de Sitter universe, and the fish analogy). The second is rather more hybridly reticulated, as in *MP*, which has progression from one textual grouping to another (from mathematics to the landscape blend), but also excursus (from mathematics to alchemy, and from landscapes to shadows). A rough scale can therefore be described, from a

⁵¹ It is known that Ryder was developing this idea further before her death in 2008, for example in a conference paper given in the same year. The elementary expansion above is unlikely to have fitted well into a wider and more comprehensive theory such as Ryder would have developed. The concept of this analysis of progression as used here was, however, inspired by discussions with Ryder, and while no publications of hers expand it in this direction, reference to her work is appropriate here.

wholly linear atemporal progression on one side to a fully matrical reticulation on another. On this scale, the key extract texts sit from *WP* to *FC*, to *MP* and finally *PW*. A computational analysis of this would be possible, describing the distance between the first and last input spaces and the complexity of the path between the two.⁵² This measure, crude as it is, can act as a reflex of the act of reading an expository text similar to Genette's important concept of order in narrative progression (1980:33ff); a linear expository progression trace would mirror a simple narrative, while a matrical one would mirror the experience of analepsis/prolepsis and external reach of narrative diegesis. Future research in this area could focus on adverbial discourse markers as excursus markers (*however, regardless, and so on*; cf Biber *et al* 1999:1086ff) in addition to reader-response work on the expository 'shape' of texts.

5.IV: CONCLUSION

This chapter has completed the core analysis of this thesis, taking the key extracts identified above and confirming the hypothesis outlined in Chapter 1 regarding popular science. Beyond this, it has offered some developments to conceptual blending theory based on this data, including a description of the acceptability of multimodality, the nature of expository progression, and the style of texts compared to new proposed sub-types of blend. These are summarized further in the following chapter, which concludes the thesis overall and points to possible new research based on this and preceding chapters.

⁵² For this description of complexity, a maximum entropy random walk algorithm, which determines the probability of traversal between two vertices, would achieve this. See further Blanco and Lioma 2007.

CHAPTER 6: CONCLUSION

6.1: SUMMARY

Using a combination of digital and cognitive-linguistic approaches, this thesis has explored the use of reification and analogy in highly abstract popular science, and has done so with the overall aim of outlining a new methodological approach to this analysis. By showing through corpus methodologies that analogical material is statistically prevalent in a corpus of texts of this type, I have aimed to avoid text-selection bias and demonstrate some the possibilities of the use of new lexicological resources in digital humanities studies. Taking this corpus-based data and using it as a selection tool for cognitive linguistic studies then showed the potential for a combination of these approaches to give both ‘top-down’ and ‘bottom-up’ perspectives on a text; by using the same corpus and similar techniques to discover highly-analogical text portions for later analysis, the thesis continued to attempt to avoid analyst selection bias. By then taking these key extracts and devising a further methodology which drew on one of the lexicological resources used earlier, the thesis grounded the cognitive-linguistic analysis as much as possible in a detailed methodology, which leaned where possible on external empirical sources as support for its interpretative approach. Finally, this analytical process was applied to the identified key extracts in order to draw out some observations on the discourse of highly-abstract popularized science texts. In this way, the thesis has offered new methodologies and perspectives on some aspects of linguistic analysis, alongside a comprehensive worked example which illustrates this analytic development. The analysis undertaken is significant in its own right, not just as an example of a successful analytic procedure, but also as a testing of often-cited assertions concerning reification, metaphor and concretization. Overall, although this thesis has moved between different, but not incompatible, schools of linguistic inquiry (distant reading,

corpus linguistics, cognitive semantics, and discourse analysis), at each stage the considerations of one has informed the use of the other, to their mutual enrichment.

6.1.1: CHAPTER 2: FROM SEMANTICS TO CORPORA

In more detail, following the introductory chapter, the first part of the thesis began with Chapter 2, which gave a comprehensive methodological justification and exposition of the thesis' proposed process of semantically analysing large bodies of text using the recently-completed database of the *Historical Thesaurus*. It was demonstrated that relatively straightforward and open-source tools can be implemented on text corpora and integrated into the *HT* database in a way which gives results which can be used for statistical analysis.

6.1.2: CHAPTER 3: FROM CORPORA TO ANALOGICAL DOMAINS

The log-likelihood statistical measure of textual relevance, when run on these texts, produced in Chapter 3 ranked lists of semantic domains which were key in the meaning construction of the discourse. These lists had to be manually pruned (see §6.ii below), and once this was done, the process had produced data of the most statistically significant semantic domains which occurred in each text. These key domains were then separated into two categories; those which were directly relevant to the text based on paratextual metadata, and those which were not. The consistent appearance in the popular science corpus of analogical domains, not relevant to the text's literal concerns, but used often in an analogical sense, proved both the existence and the extent of significant analogical content in this genre.

6.1.3: CHAPTER 4: FROM ANALOGICAL DOMAINS TO CONCEPTUAL BLENDS

With these empirical foundations, Chapter 4 continued the theme of aiming at a certain amount of externally-supported methodological rigour within the

'bottom-up' approach. Accordingly, it outlined a technique for identifying representative samples of analogical text with a minimum of analyst intuition, using small textual clusters and graphs of the presence of representative analogical lexical items. These samples were then passed to the cognitive analysis which followed. Before this, however, it was necessary to include a detailed exposition of Conceptual Integration Theory as an explicit codification of the version of CIT used in this thesis. Chapter 4 then concluded with a lengthy and detailed explanation of the adaptation of empirical and corpus methods to support and ground CIT analysis, informed by the earlier development of methodologies for the study of analogy on a wider scale. This included proposals for the use of *HT* lexicological data in the analysis alongside pragmatic theories of textual relevance, alongside a somewhat more procedurally-detailed system of undertaking CIT analyses. This was illustrated by the comprehensive worked example of a further short extract from the highly-abstract popular science corpus, illustrating the procedure in some detail.

6.1.4: CHAPTER 5: FROM CONCEPTUAL BLENDS TO DISCOURSE

Finally, Chapter 5 took the example corpus provided at the start of Chapter 4 and, using the methodology explicated at the end of that same chapter, produced a detailed analysis of the example corpus, with many of the analytic details in appendices for the sake of brevity. Chapter 5 therefore finished the thesis with the conclusions which can be drawn from such analyses, completing the methodological chain of reasoning from wide-scale corpora to narrow-focus semantics, and providing information about the nature of highly-abstract popular science as a genre.

6.1.5: OVERALL: FROM SEMANTICS TO DISCOURSE

Overall, the thesis makes eight key contributions:

1. The application of *HT* data to automated textual tagging, for the first time.
2. The empirical semantic proof of significant analogical content in a corpus, due to the division between analogical and subject-relevant content as shown by the *HT* significant domain identification process.
3. The corpus-based identification of a class of lexemes here called analogical markers, which lexically indicate that analogical content follows.
4. The demonstration of a visual-empirical technique for identifying key analogical extracts in a text which has gone through the procedure in 2 above.
5. The integration of various data-driven technical heuristics into Conceptual Integration Theory for the purpose of detailed and reproducible cognitive analysis of texts, including a heuristic based on *HT* data.
6. The proof of the use of reification as a discourse strategy in this genre, with data on the relative frequency of concrete to abstract domains in the sample highly-analogical popular science text corpus.
7. The schematization of blended spaces into original/redefinition types, as well as the concept of a pre-enacted blend.
8. Tentative steps towards an analysis of expository progression in the style of narrative analysis, using blending diagrams to display this progression, and in the schematization of texts towards a cline from expository linearity to reticulated matricularity.

Beyond this, the thesis also aims to point the way towards a more empirically-focused cognitive linguistics, by acting as a proof-of-concept demonstrating

techniques and approaches that can easily be integrated into a cognitive analytic workflow to minimize analyst intervention and intuition.

6.II: EFFECTIVENESS AND SHORTCOMINGS

With that aim in mind, at each point in the above process, weak links were highlighted in the thesis' ideal (but unattainable) goal of having as much analysis done routinely and procedurally, following an explicit discovery procedure, such as a computer would be able to follow. While such areas were to be expected, the thesis minimized them where possible, resulting in weak links only existing in areas which point to necessary lacunae in the textual digital humanities. These include the following:

1. The manual pruning of data in §3.ii, where the problem of polysemy caused a number of key domains to be misidentified, requiring a significant level of intervention in the data. Both below in §6.iii and above in §2.iii, the case for a semantic tagger based on the *HT* has been made, and the existence of such a tool would easily solve this problem.
2. The visual nature of the key extract identification in §4.ii, which used a graph to aid an analyst to identify these extracts. As with any phenomenon displayed on a graph, there is a solely numerical algorithm which could have provided an answer to this puzzle straightforwardly. However, as the graph allowed an analyst to easily examine three dimensions of data simultaneously (clustering of extract chunks on the x axis, total height of marker occurrences, and proportionality of marker categories), this was considered preferable to a statistical analysis for reasons of brevity.
3. Deciding which *HT* category a particular domain enactor referenced in §4.v.4.b. This would again be solved by the existence of a semantic tagger, as in I above.

4. Easily identifying abstract and concrete domains in §5.iii.2. This will, as signalled in that section, be made easier by the University of Glasgow/ AHRC research project Mapping Metaphor,⁵³ which aims to investigate metaphoricity in English using the *HT*, and will address the nature of domain mappings, necessitating engagement with the concrete and abstract as encoded in the *HT* ontology.

For each of these, further developments are necessary to gap-fill the procedure outlined here and therefore contribute to the explanatory and descriptive potential of the digital humanities and to empirical cognitive linguistics. Certain of these developments would constitute entire theses in themselves, in various fields, and so were not addressed in this present work – although it is hoped that the descriptions above, and this present section, will contribute towards these goals.

6.III: FUTURE WORK

With the limitations in §6.ii above in mind, future work in this area should centre around those objections. In particular, a semantic tagger with the power and granularity of the *HT* (not to mention its possible historical dimension) would be a powerful boost to the abilities of the digital humanities, particularly in the area of corpus linguistics, but also with significant impact potential in many other fields, including data mining, ontological search, and relevance determination. The development of this, along the lines of those proposed at the end of §2.iii above is, at present, a high priority for *HT* researchers.

The second main area for further work is in the expansion of the contributions listed in §6.i.5, in particular the final point about textual progression traces, which is currently highly tentative and theoretical without a much larger

⁵³ Scheduled to start in January 2012, under Principal Investigator Wendy Anderson, with Co-Investigators Christian Kay and Carole Hough. See also Alexander and Kay 2011a [2010] for a pilot of the methodology of this project.

corpus. The schematization of blended spaces could also be improved, and the concretization work could be expanded into a much wider corpus of popular science, once work on improving the automation of the entire procedure is complete (for which, see §6.ii).

The third, and most minor, area is more textual; the arch-blend in the Sautoy *The Music of the Primes* extract above is incomplete due to the limitations of space in this thesis, and the requirement to have all extracts of roughly equivalent length. Completion of this extract would test the hypothesis in §5.iii.2 regarding this text, and would also serve as a testbed for exploring the limits of the diagrammatic method of displaying CIT networks.

6.IV: *ENVOI*

In one of the few linguistic pronouncements of Noam Chomsky with which the present author fully agrees, the father of modern generative grammar stated in interview that ‘If somebody today is teaching what they taught ten years ago, then either their subject is dead or they are dead’ (Pacitti 2000:21). Following this, it is most sincerely hoped that the contributions outlined in this present work will form part of a chain of work which, in the best traditions of the digital humanities, aim to ask questions which were not previously known to be asked. This is the natural consequence of a field engaged with developing and implementing constantly-evolving tools and methodologies.

Therefore, while forming a complete work in itself, this thesis aims to be part of such a chain – leading, where possible, to future work and methodologies which would make the sort of analysis painstakingly outlined and implemented in previous chapters automatic, rigorous, and reliable. At such a point, the questions which a linguistic analyst could fruitfully address would increase considerably, and to great benefit.

BIBLIOGRAPHY

- Abbott, Edwin A. 1884. *Flatland: A Romance in Many Dimensions*. London: Seely and Co. Facsimile version available from <http://www.archive.org/details/flatlandromanceoooabbouoft> (5 March 2009).
- Alexander, Marc. 2006. *Cognitive-Linguistic Manipulation and Persuasion in Agatha Christie*. University of Glasgow: M.Phil. thesis.
- Alexander, Marc. 2010a. *Fizzing, Eximious, and All Wool and a Yard Wide: Conceptual Search Across Integrated Resources*. Conference paper at *Text-Mining in the Digital Humanities*, University of Lancaster.
- Alexander, Marc. 2010b. *My Body Part Verbed: Transitivity Analysis, Enroller and the Historical Thesaurus of English*. Conference paper at *Poetics and Linguistics Association: The Language of Landscapes*, University of Genoa, Italy.
- Alexander, Marc. 2011a. 'The Various Forms of Civilization Arranged in Chronological Strata': *Manipulating the Historical Thesaurus of the OED*. In Michael Adams (ed.), *'Cunning passages, contrived corridors': Unexpected Essays in the History of Lexicography*. Monza: Polimetrica.
- Alexander, Marc. 2011b. *Research Implications of the Historical Thesaurus*. Conference paper at *Historical Semantics, Etymology and Lexicography: A Meeting of the Philological Society*, University of Glasgow.
- Alexander, Marc & Christian Kay. 2011a [2010]. *Mapping Metaphors Across Time with the Historical Thesaurus*. Conference paper at *Helsinki Corpus Festival: The Past, Present, and Future of English Historical Corpora*, University of Helsinki, Finland. Based on an earlier paper at *The 3rd UK Cognitive Linguistics Conference*, University of Hertfordshire.
- Alexander, Marc & Christian Kay. 2011b. *A Digital Approach to Sound Symbolism in English: Evidence from the Historical Thesaurus*. Conference paper at *Language in Glasgow*, University of Glasgow.
- Alexander, Marc & Andrew Struan. In press, 2012. 'In countries so unciviliz'd as those?': *Notions of Civility and the British Experience of the World*. In Martin Farr & Xavier Guégan (eds.), *Experiencing Imperialism: Interdisciplinary and Transnational Perspectives on the Colonial and Post-Colonial British*. London: Palgrave Macmillan.

- Alexander, Marc & Catherine Wild. 2010. *Men, Women, and Children in the Historical Thesaurus: A Case Study*. In *OED Online*. Oxford: OUP. <http://www.oed.com> (15 August 2011).
- Algeo, John. 1990. The Emperor's New Clothes: The Second Edition of the Society's Dictionary. *Transactions of the Philological Society* 88(2). 131-150.
- Anderson, Jean, Marc Alexander, Johanna Green, Muhammad Sulman Sarwar, & Richard Sinnott. 2011a. *Enroller: A Digital Research Platform for English and Scots*. *Digital Humanities 2011*. <http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-271.xml;query=&brand=default> (18 September 2011).
- Anderson, Jean, Marc Alexander, Christian Kay, & Muhammad Sulman Sarwar. 2011b. Data Mining and Search Enhancements using the *Historical Thesaurus of English*. *Proceedings of the 1st International Workshop on Search and Mining Entity-Relationship Data*. <http://dl.acm.org/citation.cfm?id=2065000> (8 November 2011).
- Archer, Dawn. 2011. Corpus Annotation – A Welcome Addition or an Interpretation Too Far? Conference plenary at *Helsinki Corpus Festival: The Past, Present, and Future of English Historical Corpora*, University of Helsinki, Finland.
- Archer, Dawn, Andrew Wilson, & Paul Rayson. 2002. Introduction to the USAS Category System. <http://ucrel.lancs.ac.uk/usas/usas%20guide.pdf> (7 November 2011).
- Archer, Dawn & Christian Kay. 2010. Out of the *darkness...*: On the colour *black*. Conference paper at *Text-Mining in the Digital Humanities*, University of Lancaster.
- Baayen, R. Harald. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.
- Baker, Paul. 2010. *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, & Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.

- Bird, Steven, Ewan Klein, & Edward Loper. 2009. *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly.
- Bod, Rens, Jennifer Hay, & Stefanie Jannedy (eds.). 2003. *Probabilistic Linguistics*. Cambridge, MA: MIT Press.
- Boroditsky, Lera & Michael Ramscar. 2002. The Roles of Body and Mind in Abstract Thought. *Psychological Science* 13. 185–188.
- Boufaden, Narjès. 2003. An Ontology-based Semantic Tagger for IE System. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. 7-14.
- Brandt, Line. 2002. *Conceptual Integration (4.0)* [Working Paper]. <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=4D82CC3EAB9E487E11E205D4342E479B?doi=10.1.1.195.2012&rep=rep1&type=pdf> (24 August 2010).
- Brandt, Line. 2010. *Language and Enunciation – A Cognitive Inquiry with Special Focus on Conceptual Integration in Semiotic Meaning Construction*. Aarhus University: Ph.D. thesis. http://www.hum.au.dk/semiotics/docs2/pdf/brandt_line_phd/Brandt_manuscript.pdf (8 March 2011).
- Brandt, Per Aage. 2005. Mental Spaces and Cognitive Semantics: A Critical Comment. *Journal of Pragmatics* 37(10). 1578-1594.
- Burchfield, Robert W. (ed.). 1972-1986. *Oxford English Dictionary Supplement*. 4 vols (vol. 1: A-G, 1972; vol. 2: H-N, 1976; vol. 3: O-Scz, 1982; vol. 4: Se-Z, 1986). Oxford: Clarendon Press.
- Butler, Christopher S. 1990. Qualifications in Science: Modal Meanings in Scientific Texts. In Walter Nash (ed.), *The Writing Scholar: Studies in Academic Discourse*. London: Sage.
- Coppel, William A. 2009. *Number Theory: An Introduction to Mathematics*. 2nd edn. London: Springer.
- Corbett, John B. 1992. *Functional Grammar and Genre Analysis: A Description of the Language of Learned and Popular Articles*. University of Glasgow: Ph.D. thesis.
- Corbett, John B. 2006. Popularisations. In Keith Brown (ed.), *Encyclopedia of Language and Linguistics*, 2nd edn. Oxford: Elsevier.
- Coulson, Seana. 2001. *Semantic Leaps: Frame-shifting and Conceptual Blending in Meaning Construction*. Cambridge: Cambridge University Press.

- Coulson, Seana & Todd Oakley. 2000. Blending Basics. *Cognitive Linguistics* 11(3-4). 175-196.
- Coulson, Seana & Todd Oakley. 2003. Metonymy and Conceptual Blending. In Klaus-Uwe Panther & Linda Thornburg (eds.), *Metonymy and Pragmatic Inferencing*, 51-79. Amsterdam: John Benjamins.
- Craigie, William A. & Charles T. Onions (eds.). 1933. *Oxford English Dictionary Supplement*. Oxford: Clarendon Press.
- Croft, William & D. Alan Cruse. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.
- Dabrowska, Ewa. 2004. *Language, Mind and Brain: Some Psychological and Neurological Constraints on Theories of Grammar*. Edinburgh: Edinburgh University Press.
- Darwin, Charles R. 1859. *On the Origin of Species*. London: John Murray.
Facsimile version available from http://darwin-online.org.uk/pdf/1859_Origin_F373.pdf (25 August 2011).
- Dancygier, Barbara. 2006. What can Blending Do for You? *Language and Literature* 15(5). 5-15.
- Dancygier, Barbara. Forthcoming. *The Language of Stories: A Cognitive Approach*. Cambridge: Cambridge University Press.
- Davidson, George W. 2004. *Roget's Thesaurus of English Words and Phrases*. London: Penguin.
- Demetriou, George & Eric S. Atwell. 2001. A Domain-Independent Semantic Tagger for the Study of Meaning Associations in English Text. *Proceedings of the Fourth International Workshop on Computational Semantics*. 67-80.
<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.8.5043> (6 October 2011).
- van Dijck, Jose. 1998. *Imagination: Popular Images of Genetics*. London: MacMillan.
- Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1). 61-74.
- Emmott, Catherine. 1997. *Narrative Comprehension: A Discourse Perspective*. Oxford: Oxford University Press.

- Emmott, Catherine. 2003. Reading for Pleasure: A Cognitive Poetic Analysis of 'Twists in the Tale' and Other Plot Reversals in Narrative Texts. In Gerard Steen & Joanna Gavins (eds.), *Cognitive Poetics in Practice*. London: Routledge.
- Emmott, Catherine & Marc Alexander. 2009/2010. Schema Theory. In Peter Hühn, John Pier, Wolf Schmid, & Jorg Schönert (eds.), *Handbook of Narratology*, 2009. Berlin: de Gruyter. Also *The Living Handbook of Narratology*, 2010. Hamburg: Hamburg University Press. <http://hup.sub.uni-hamburg.de/lhn/index.php/Schemata> (31 February 2011).
- Emmott, Catherine, Anthony J. Sanford, & Lorna I. Morrow. 2006. Capturing the Attention of Readers? Stylistic and Psychological Perspectives on the Use and Effect of Text Fragmentation in Narratives. *Journal of Literary Semantics* 35(1). 1-30.
- Evans, Vyvyan. 2009. *How Words Mean*. Oxford: Oxford University Press.
- Evans, Vyvyan & Melanie Green. 2006. *Cognitive Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
- Evans, Vyvyan & Stephanie Pourcel. 2009. *New Directions in Cognitive Linguistics*. Amsterdam: John Benjamins.
- Evans, Vyvyan, Benjamin K. Bergen, & Jorg Zinken. 2007. The Cognitive Linguistics Enterprise: An Overview. In Vyvyan Evans, Benjamin K. Bergen, & Jorg Zinken (eds.), *The Cognitive Linguistics Reader*. London: Equinox.
- Everett, Dan. 2002. Falsifiability vs. Usefulness. *Linguist List* 13.1279. <http://linguistlist.org/issues/13/13-1279.html> (3 August 2011).
- Fauconnier, Gilles. 1985. *Mental Spaces*. Cambridge: Cambridge University Press.
- Fauconnier, Gilles & Mark Turner. 1998. Conceptual Integration Networks. *Cognitive Science* 22(2). 133-87.
- Fauconnier, Gilles & Mark Turner. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.
- Fauconnier, Gilles & Eve Sweetser (eds.). 1996. *Spaces, Worlds and Grammar*. Chicago: University of Chicago Press.
- Fauconnier, Gilles & George Lakoff. 2010. *On Metaphor and Blending*. <http://www.cogsci.ucsd.edu/~coulson/spaces/GG-final-I.pdf> (5 August 2011).

- Feldman, Jerome A. 2006. *From Molecule to Metaphor: A Neural Theory of Language*. Cambridge, MA: MIT Press.
- Fowler, Henry W. & Francis G. Fowler. 1911. *The Concise Oxford Dictionary of Current English*. Oxford: Clarendon Press.
- Garretson, Gregory & Mary C. O'Connor. 2007. Between the Humanist and the Modernist: Semi-automated Analysis of Linguistic Corpora. In Eileen Fitzpatrick (ed.), *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*. Amsterdam: Rodopi. 87-106.
- Gavins, Joanna. 2007. *Text World Theory: An Introduction*. Edinburgh: Edinburgh University Press.
- Geeraerts, Dirk. 2006. A Rough Guide to Cognitive Linguistics. In Dirk Geeraerts (ed.), *Cognitive Linguistics: Basic Readings*. Berlin: de Gruyter. 1-28.
- Geeraerts, Dirk & Herbert Cuyckens (eds.). 2007a. *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford University Press.
- Geeraerts, Dirk & Herbert Cuyckens. 2007b. Introducing Cognitive Linguistics. In Geeraerts & Cuyckens 2007a. 3-24.
- Genette, Gerard. 1980. *Narrative Discourse: An Essay in Method*. Trans. Jane E. Lewin. Ithaca, NY: Cornell University Press.
- Gentner, Dedre & Brian Bowdle. 2008. Metaphor as Structure-mapping. In Ray Gibbs (ed.) *The Cambridge Handbook of Metaphor and Thought*. Cambridge: Cambridge University Press. 109-128.
- Gibbs, Raymond W. 1990. Psycholinguistic Studies on the Conceptual Basis of Idiomaticity. *Cognitive Linguistics* 1. 417-462.
- Gibbs, Raymond W. 2000. Making Good Psychology out of Blending Theory. *Cognitive Linguistics* 11(3-4). 347-358.
- Gibbs, Raymond W. 2006a. Introspection and Cognitive Linguistics: Should We Trust Our Own Intuitions? *Annual Review of Cognitive Linguistics* 4(1). 135-151.
- Gibbs, Raymond W. 2006b. Metaphor Interpretation as Embodied Simulation. *Mind and Language* 21. 434 - 458.
- Grady, Joseph. 2005. Primary Metaphors as Inputs to Conceptual Integration. *Journal of Pragmatics* 37(10). 1595-1614.

- Grady, Joseph, Todd Oakley, & Seana Coulson. 1999. Conceptual Blending and Metaphor. In Gerard Steen & Raymond Gibbs (eds.), *Metaphor in Cognitive Linguistics*, 101-124. Amsterdam: John Benjamins.
- Greene, Brian R. 2004. *The Fabric of the Cosmos: Space, Time and the Texture of Reality*. Alfred A Knopf: New York.
- Gries, Stefan T.H. 2009. *Qualitative Corpus Linguistics with R: A Practical Introduction*. London: Routledge.
- Gries, Stefan T.H. & Anatol Stefanowitsch (eds.). 2006. *Corpus-based Approaches to Metaphor and Metonymy*. Berlin: de Gruyter.
- Gries, Stefan T.H., Beate Hampe, & Doris Schonefeld. 2005. Converging Evidence: Bringing Together Experimental and Corpus Data on the Association of Verbs and Constructions. *Cognitive Linguistics* 16(4). 635–676.
- Grondelaers, Stefan, Dirk Geeraerts, & Dirk Speelman. 2007. A Case for a Cognitive Corpus Linguistics. In Gonzales-Marques, Monica, Irene Mittelberg, Seana Coulson, & Michael J. Spivey (eds.), *Methods in Cognitive Linguistics*. Amsterdam: John Benjamins. 149-169.
- Halliday, Michael A.K. and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.
- Hardy, Donald E. 2005. Towards a Stylistic Typology of Narrative Gaps: Knowledge Gapping in Flannery O'Connor's Fiction. *Language and Literature* 14(5). 363-376.
- Hofstadter, Douglas R. 1979. *Gödel, Escher, Bach: An Eternal Golden Braid*. 2nd ed 1999. New York: Basic Books.
- Hudson, Richard. 2007. *Language Networks. The New Word Grammar*. Oxford: Oxford University Press.
- Hüllen, Werner. 2004. *A History of Roget's Thesaurus: Origins, Development, and Design*. Oxford: Oxford University Press.
- Hüllen, Werner. 2009. *Networks and Knowledge in Roget's Thesaurus*. Oxford: Oxford University Press.
- Israel, Michael, Jennifer Riddle Harding, & Vera Tobin. 2005. On Simile. In Suzanne Kemmer & Michel Achard (eds.), *Language, Culture and Mind*. Stanford: CSLI Publications. 123-135.

- Jeffries, Lesley. 2000. *Don't Throw Out the Baby with the Bathwater: In Defence of Theoretical Eclecticism in Stylistics*. PALA Occasional Papers 12. <http://www.pala.ac.uk/resources/op/Paper12.pdf> (2 January 2007).
- Johnson, Anthony, Lisa Lena Opas-Hänninen, & Ilkka Juuso. 2010. Stitches in Time and Switches in Text: Diana Gabaldon and the *Historical Thesaurus of the Oxford English Dictionary*. Conference paper at *Poetics and Linguistics Association: The Language of Landscapes*, University of Genoa, Italy.
- Juuso, Ilkka, Lisa Lena Opas-Hänninen, Anthony Johnson, & Tapio Seppänen. 2011. *The Time Machine: Capturing Worlds across Time in Texts*. Digital Humanities 2011. <http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-113.xml;brand=default> (18 September 2011).
- Kaku, Michio. 2005. *Parallel Worlds; A Journey Through Creation, Higher Dimensions, and the Future of the Cosmos*. New York: Doubleday.
- Kay, Christian, Jane Roberts, Michael Samuels, & Iréne Wotherspoon (eds.). 2009. *Historical Thesaurus of the Oxford English Dictionary*. Oxford: Oxford University Press.
- Kay, Christian & Marc Alexander. 2010. Colour in the *Historical Thesaurus of the Oxford English Dictionary*. Conference paper at *The Fifth International Conference on Historical Lexicography and Lexicology*, University of Oxford.
- Kemmer, Suzanne. 2010. *About Cognitive Linguistics: Historical Background*. <http://www.cogling.org/cl.shtml> (11 December 2010).
- Kress, Gunther & Theo van Leeuwen. 2001. *Multimodal Discourse: The Modes and Media of Contemporary Communication*. London: Arnold.
- Kress, Gunther & Theo van Leeuwen. 2006. *Reading Images: The Grammar of Visual Design*. 2nd edn. London: Routledge.
- Knorr-Cetina, Karin and Michael J. Mulkay (eds.). 1983. *Science Observed*. London: Sage.
- Kövecses, Zoltán. 2002. *Metaphor: A Practical Introduction*. Oxford/New York: Oxford University Press.
- Lakoff, George. 1987. *Women, Fire and Dangerous Things*. Chicago: University of Chicago Press.
- Lakoff, George. 1990. The Invariance Hypothesis: Is Abstract Reason Based on Image-Schemas? *Cognitive Linguistics* 1. 39-74.

- Lakoff, George. 1993. The Contemporary Theory of Metaphor. In Andrew Ortony (ed.), *Metaphor and Thought*. 2nd edn. Cambridge: Cambridge University Press. 202-251.
- Lakoff, George & Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.
- Lakoff, George & Mark Johnson. 1999. *Philosophy in the Flesh*. New York: Basic Books.
- Lakoff, George & Mark Johnson. 2003. Afterword. In Lakoff, George & Mark Johnson, *Metaphors We Live By*. 2nd edn.
- Lakoff, George & Mark Turner. 1989. *More than Cool Reason: A Field Guide to Poetic Metaphor*. Chicago: University of Chicago Press.
- Langacker, Ronald. 2002. *Concept, Image and Symbol: The Cognitive Basis of Grammar*. Berlin: de Gruyter.
- Larose, Daniel. 2006. *Data Mining Methods and Models*. Hoboken, NJ: Wiley.
- Leech, Geoffrey & Mick Short. 1981. *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. London: Longman.
- Leech, Geoffrey & Jan Svartvik. 2002. *A Communicative Grammar of English*. 3rd edn. London: Longman.
- van Leeuwen, Theo & Carey Jewitt (eds.). 2001. *The Handbook of Visual Analysis*. London: Sage.
- Louw, Bill. 1993. Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. In Baker, Mona, Gill Francis, & Elena Tognini-Bonelli (eds.), *Text and Technology*. Amsterdam: John Benjamins.
- Low, Graham, Alice Deignan, & Zazie Todd. 2010. *Researching and Applying Metaphor in the Real World*. Amsterdam: John Benjamins.
- Mandelblit, Nili. 2000. The Grammatical Marking of Conceptual Integration: From Syntax to Morphology. *Cognitive Linguistics* 11(3-4). 197-252.
- Markman, Arthur. 1999. *Knowledge Representation*. Mahwah, NJ: Erlbaum.
- Mazur, Barry. 2004. *Imagining Numbers: Particularly the Square Root of Minus Fifteen*. Harmondsworth: Penguin.
- McArthur, Tom. 1981. *The Longman Lexicon of Contemporary English*. London: Longman.

- McIntyre, Dan & Dawn Archer. 2010. A Corpus-based Approach to Mind Style. *Journal of Literary Semantics* 39(2). 167-182.
- McIntyre, Dan & Brian Walker. 2010. How can Corpora be Used to Explore the Language of Poetry and Drama? In Anne O’Keeffe & Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge. 516-530.
- Mendoza-Denton, Norma, Jennifer Hay, & Stefanie Jannedy. 2003. Probabilistic Sociolinguistics: Beyond Variable Rules. In Bod *et al* 2003.
- Meyer, Charles F. 2002. Pseudo-Titles in the Press Genre of Various Components of the International Corpus of English. In Randi Reppen, Susan M. Fitzmaurice, & Douglas Biber (eds.), *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins. 147-166.
- Moretti, Franco. 2000. Conjectures on World Literature. *New Left Review* 1. <http://newleftreview.org/A2094> (1 October 2007).
- Moretti, Franco. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.
- Murray, James A. H., Henry Bradley, William A. Craigie, & Charles T. Onions (eds.). 1884-1928. *A New English Dictionary on Historical Principles*. [=Oxford English Dictionary, 1st edn], 10 vols/128 fascicles. Oxford: Clarendon Press.
- Myers, Greg. 1990a. *Writing Biology: Texts in the Social Construction of Scientific Knowledge*. Madison, WI: University of Wisconsin Press.
- Myers, Greg. 1990b. Making a Discovery: Narratives of Split Genes. In Walter Nash (ed.), *The Writing Scholar: Studies in Academic Discourse*. London: Sage. 102-126.
- Myers, Greg. 1994. Narratives of Science and Nature in Popularizing Molecular Genetics. In Malcolm Coulthard (ed.), *Advances in Written Text Analysis*. London: Routledge. 179-190.
- Neiman, Adam. 2000. *The Popularisation of Physics: Boundaries of Authority and the Visual Culture of Science*. University of the West of England: Ph.D. thesis.
- Oakley, Todd. 1998. Conceptual Blending, Narrative Discourse and Rhetoric. *Cognitive Linguistics* 9(4). 321-360.
- Pacitti, Domenico. 2000. A New Twist in an Old Turn of Phrase [Interview with Noam Chomsky]. *Times Higher Education Supplement*, 24.03.2000.

- Pederson, Eric. 2007. Cognitive Linguistics and Linguistic Relativity. In Geeraerts & Cuyckens 2007a. 1012-1044.
- Proffitt, Michael & John A. Simpson (eds.). 1997. *Oxford English Dictionary Additions Series*, vol. 3. Oxford: Clarendon Press.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, & Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
- Randall, Lisa. 2005. *Warped Passages: Unraveling the Mysteries of the Universe's Hidden Dimensions*. New York: Harper Perennial.
- Rayson, Paul. 2008. From Key Words to Key Semantic Domains. *International Journal of Corpus Linguistics* 13.4. 519-549.
- Rayson, Paul. 2009. *Wmatrix: a web-based corpus processing environment*. Lancaster University Computing Department. <http://ucrel.lancs.ac.uk/wmatrix/>
- Rayson, Paul, Damon Berridge, & Brian Francis. 2004. Extending the Cochran Rule for the Comparison of Word Frequencies between Corpora. *7th International Conference on Statistical analysis of textual data*. <http://eprints.comp.lancs.ac.uk/893/> (1 April 2010).
- Reisberg, Daniel. 1997. *Cognition: Exploring the Science of the Mind*. New York: Norton.
- Reppen, Randi, Susan Fitzmaurice, & Douglas Biber (eds.). 2002. *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins.
- Roberts, Jane & Christian Kay, with Lynne Grundy. 2000 [1995]. *A Thesaurus of Old English*. Amsterdam: Rodopi.
- Roget, Peter Mark. 1852. *Thesaurus of English Words and Phrases*. London: Longman, Brown, Green and Longman.
- Rozenfeld, Boris Abramovich. 1988. *A History of Non-Euclidean Geometry: Evolution of the Concept of a Geometric Space*. London: Springer.
- Ryan, Marie-Laure. 1991. *Possible Worlds, Artificial Intelligence, and Narrative Theory*. Bloomington: Indiana University Press.
- Ryder, Mary Ellen. 2003. I Met Myself Coming and Going: Co(?)-Referential Noun Phrases and Point of View in Time Travel Stories. *Language and Literature* 12(3). 213-232.

- Sanford, Anthony J. & Simon C. Garrod. 1981. *Understanding Written Language: Explorations of Comprehension Beyond the Sentence*. Chichester: John Wiley & Sons.
- Sanford, Anthony J. & Simon C. Garrod. 1998. The Role of Scenario Mapping in Text Comprehension. *Discourse Processes* 26(2&3). 159-190.
- du Sautoy, Marcus. 2003. *The Music of the Primes: Why an Unsolved Problem in Mathematics Matters*. London: Harper Perennial.
- Schank, Roger C. & Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding*. Hillsdale, NJ: Lawrence Erlbaum.
- Schilder, Frank and Christopher Habel. 2001. From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages. *Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing*.
- Schreibman, Susan, Ray Siemens, & John Unsworth. 2004. The Digital Humanities and Humanities Computing: An Introduction. In Susan Schreibman, Ray Siemens, & John Unsworth (eds.), *A Companion to Digital Humanities*. Oxford: Blackwell. Available online at <http://www.digitalhumanities.org/companion/> (1 May 2009).
- Scott, Mike. 2008. *Wordsmith Tools Version 5*. Liverpool: Lexical Analysis Software Ltd.
- Semino, Elena. 2008. *Metaphor in Discourse*. Cambridge: Cambridge University Press.
- Short, Mick, Donald C. Freeman, Willie van Peer, & Paul Simpson. 1998. Stylistics, Criticism and Mythrepresentation Again: Squaring the Circle with Ray Mackay's Subjective Solution for All Problems. *Language and Literature* 7(1). 39-50.
- Simpson, John A. 2007. *March 2007 Update*. <http://www.oed.com/public/update0703/march-2007-update#rev> (2 May 2011).
- Simpson, John A. & Edmund S. C. Weiner *et al* (eds.). 1989. *Oxford English Dictionary*. 2nd edn. [=Murray *et al* 1989.] Compiled by John A. Simpson & Edmund S. C. Weiner from Murray *et al* 1884, Craigie & Onions 1933, Burchfield 1972-1986, and additional corrections edited by John A. Simpson & Edmund S. C. Weiner. 20 vols. Oxford: Clarendon Press.
- Simpson, John A. & Edmund S. C. Weiner (eds.). 1993. *Oxford English Dictionary Additions Series*, vols 1-2. Oxford: Clarendon Press.

- Simpson, John A. & Edmund S. C. Weiner *et al* (eds.). 2000-. *Oxford English Dictionary Online*. 3rd edn. Revised by John A. Simpson and others from Murray *et al* 1884 and Simpson & Weiner *et al* 1989. Online only at <http://www.oed.com> (9 October 2011).
- Sinclair, John M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, John M. 2004. *Trust The Text: Language, Corpus and Discourse*. London: Routledge.
- Sinclair, John M. 2007. The Exploitation of Meaning: Literary Text and Local Grammars. In Isil Bas & Donald C. Freeman (eds.), *Challenging the Boundaries*. Amsterdam: Rodopi.
- Soanes, Catherine & Angus Stevenson. 2009. *The Concise Oxford English Dictionary*. 11th edn, 4th rev, revised from Fowler & Fowler 1911. Oxford: Oxford University Press.
- Smith, Carlota S. 2003. *Modes of Discourse: The Local Structure of Texts*. Cambridge: Cambridge University Press.
- Sperber, Dan & Deirdre Wilson. 1995. *Relevance: Communication and Cognition*. 2nd edn. Oxford: Blackwell.
- Sperber, Dan & Deirdre Wilson. 2008. A Deflationary Account of Metaphors. In Ray Gibbs (ed.) *The Cambridge Handbook of Metaphor and Thought*. Cambridge: Cambridge University Press. 84-108.
- Steen, Gerard. 2002. Identifying Metaphor in Language: A Cognitive Approach. *Style* 36. 386-407.
- Steen, Gerard. 2007. *Finding Metaphor in Grammar and Usage: A Methodological Analysis*. Amsterdam: John Benjamins.
- Stenning, Keith. 2002. *Seeing Reason: Image and Language in Learning to Think*. Oxford: Oxford University Press.
- Stockwell, Peter. 1992. 'Do Androids Dream of Electric Sheep?' Isomorphic Relations in Reading Science Fiction. *Language and Literature* 1(2). 79-99.
- Stockwell, Peter. 2000. *The Poetics of Science Fiction*. London: Longman.
- Stockwell, Peter. 2006. Schema Theory: Stylistic Applications. In Keith Brown (ed.), *Encyclopedia of Language and Linguistics*, 2nd edn. Oxford: Elsevier.

- Stoppard, Tom. 1967. *Rosencrantz and Guildenstern are Dead*. London: Faber and Faber.
- Stoppard, Tom. 1972. *Jumpers*. London: Faber and Faber.
- Stoppard, Tom. 1993. *Arcadia*. London: Faber and Faber.
- Sykes, John. 1982. *The Concise Oxford Dictionary of Current English*. 7th edn, revised from Fowler & Fowler 1911. Oxford: Clarendon Press.
- Talmy, Leonard. 2000. *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press.
- Talmy, Leonard. 2007. Foreword. In Gonzales-Marques, Monica, Irene Mittelberg, Seana Coulson, & Michael J. Spivey (eds.), *Methods in Cognitive Linguistics*. Amsterdam: John Benjamins.
- Tendahl, Markus. 2009. *A Hybrid Theory of Metaphor: Relevance Theory and Cognitive Linguistics*. Basingstoke: Palgrave.
- Tendahl, Markus & Raymond W. Gibbs. 2008. Complementary Perspectives on Metaphor: Cognitive Linguistics and Relevance Theory. *Journal of Pragmatics* 40(11). 1823-1864.
- Toolan, Michael. 2004. Graded Expectations: On the Textual and Structural Shaping of Readers' Experience. In John Pier (ed.), *The Dynamics of Narrative Form: Studies in Anglo-American Narratology*. Berlin: de Gruyter. 215-237.
- Turner, Mark. 2001. *Cognitive Dimensions of Social Science: The Way We Think About Politics, Economics, Law, and Society*. Oxford: Oxford University Press.
- Turner, Mark. 2007. *What are We? The Convergence of Self and Communications Technology (Draft)*. <http://marktturner.org/WhatAreWeTurnerDraft.pdf> (1 November 2007).
- Ungerer, Friedrich & Hans-Jorg Schmid. 2006. *An Introduction To Cognitive Linguistics*. 2nd edn. Harlow: Longman.
- Veale, Tony & Diarmuid O'Donoghue. 2000. Computation and Blending. *Cognitive Linguistics*, 11(3-4). 253-282.
- Werlich, Egon. 1976. *A Text Grammar of English*. Heidelberg: Quelle and Meyer.
- Werth, Paul. 1999. *Text Worlds: Representing Conceptual Space in Discourse*. London: Longman.

- Widdowson, Henry G. 2004. *Text, Context, Pretext: Critical Issues in Discourse Analysis*. Oxford: Blackwell.
- Wheatley, Grayson H. 1991. Constructivist Perspectives on Science and Mathematics Learning. *Science Education* 75(1). 9-21.
- Wilson, Andrew. 1993. Towards an Integration of Content Analysis and Discourse Analysis: The Automatic Linkage of Key Relations in Text. *UCREL Technical Paper* 3. <http://ucrel.lancs.ac.uk/papers/techpaper/vol3.pdf> (5 August 2009).
- Wilson, Deirdrie & Dan Sperber. 2002. Relevance Theory. *UCL Working Papers in Linguistics* 14. 249-287.
- Wilson, Deirdrie & Dan Sperber. 2004. Relevance Theory. In Lawrence R. Horn & Gregory L. Ward (eds.) *The Handbook of Pragmatics*. Oxford: Blackwell. 607-632.
- Yus, Francisco. 2009. Relevance Theory. In Jacob L. Mey (ed.), *Concise Encyclopedia of Pragmatics*. Oxford: Elsevier. 854-861.
- Zunshine, Lisa. 2006. *Why We Read Fiction: Theory of Mind and the Novel*. Columbus, OH: Ohio State University Press.

APPENDICES

You see! They wrote – they scribbled – they put it on paper. It was their employment. Their diversion. Paper is what they had. And there'll be more. There is always more.

Stoppard 1993:I.iv

APPENDIX I: PYTHON PROGRAMS AND SQL QUERIES

WORDLIST.PY

```
# WordList.py
# Author: Marc Alexander (marc@seldomend.com)
# Version: 1.9
import nltk
import re
import string
# Requires a plain text input file
raw = open('/coppel.txt').read()
output = open('/coppel.csv', 'w')
txt = raw.lower()
for punct in string.punctuation:
    txt = txt.replace(punct, ' ')
re.sub(r'\r+', ' ',txt)
stopwords = nltk.corpus.stopwords.words('english')
words = nltk.word_tokenize(txt)
filtered = [f for f in words if f.lower() not in stopwords]
wnl = nltk.WordNetLemmatizer()
lemmas = [wnl.lemmatize(t) for t in filtered]
fd = nltk.FreqDist(lemmas)
for word in fd:
    output.write(word + ',' + str(fd[word]) + '\n')
```

WORDLIST.PY: ANNOTATED

```
# WordList.py
# Author: Marc Alexander (marc@seldomend.com)
# Version: 1.9
```

This is the file heading, giving its name, origin and a version number (indicating how many times it has been revised). The # in front of the line marks the text as a comment,

```
import nltk
import re
import string
```

These three statements tell the interpreter that this program uses the NLTK toolkit module (import nltk), Python's Regular Expressions module (import re), and the Python strings module (import string) and that the interpreter needs to load those modules to understand some of the later commands.

```
# Requires a plain text input file
raw = open('/coppel.txt').read()
```

This creates a particular variable – a shortcut name for anything – that the program can access called 'raw'. It then tells the interpreter that 'raw' consists of all the text in the file coppel.txt, and so the interpreter should read the file and keep it in memory. The comment informs a reader that the input should be a text file.

```
output = open('/coppel.csv', 'w')
```

Here the interpreter is being told to create a blank file called 'coppel.csv' for writing the program's output to. The CSV extension stands for comma separated values, and is a useful way of representing text data in a spreadsheet-like way.

```
txt = raw.lower()
```

This statement creates a new shortcut called 'txt', and tells the interpreter to make 'txt' consist of all the contents of 'raw' above, but all in lowercase.

```
for punct in string.punctuation:
    txt = txt.replace(punct, ' ')
```

This uses the string module which Python provides to replace the old 'txt' above with a new 'txt', this one having all the punctuation in it (a pre-existing list defined in the string module and called 'punct') replaced by a space. Thus 'txt' is now the text of the input book, all in lower case and with no punctuation.

```
re.sub(r'\r+', ' ', txt)
```

This uses the Python Regular Expressions module to replace any end-of-line returns (\r+) in 'txt' with a space. This gets rid of end-of-line returns in any texts which include them.

```
stopwords = nltk.corpus.stopwords.words('english')
```

The predefined NLTK stopwords corpus contains those function words in English which corpus linguists generally prefer to ignore in texts. This statement takes that corpus and gives it the shorthand 'stopwords'.

```
words = nltk.word_tokenize(txt)
```

Here, the NLTK tokenizer is run on 'txt' and the results saved as the variable 'words'. This uses the Penn Treebank method, which handily splits contractions. Tokenization is here the process of taking a long string of text (which is now what 'txt' consists of) and turning it into a list of individual tokens, which the interpreter can do things with. Existing automatic tokenizers are not ideal for our purposes in this thesis, which is why the above methods of removing punctuation and line-returns were used.

```
filtered = [f for f in words if f.lower() not in stopwords]
```

This statement takes the list of words in the text created above and called 'words' and removes all the occurrences of any stopword in them. It returns another list of words which are not in the stopword list, called 'filtered'. The statement uses *f* to refer to anything in the list, and can be read as 'get the group of all *f*s so that *f* is in "words" but not in "stopwords"'. The additional '*f.lower()*' part simply lower-cases what is being looked-up in 'stopwords'.

```
wnl = nltk.WordNetLemmatizer()
```

Here the WordNet lemmatizer from NLTK is given the shorthand name of 'wnl'. This lemmatizer, in contrast to the Porter or Lancaster ones, takes words and finds their root lemma using the WordNet database (*cat* to *cats*, *running* to *run*, etc). While the WordNet Lemmatizer is not perfect, it is less 'aggressive' and therefore less noise-prone than the other rule-based ones.

```
lemmas = [wnl.lemmatize(t) for t in filtered]
```

This creates a list called 'lemmas', which consists of the lemmatized contents of 'filtered' (the part in square brackets can be read as 'lemmatize all the *t*s using "wnl", and what I mean by all the *t*s is the contents of "filtered"').

```
fd = nltk.FreqDist(lemmas)
```

Here the list of lemmas in 'lemmas' is put into a frequency distribution table – that is, they are counted and the number of times each occurs is stored alongside the word itself. This table is called 'fd'.

```
for word in fd:  
    output.write(word + ',' + str(fd[word]) + '\n')
```

Finally, we take everything in 'fd' and output it to the file called 'output' created at the start of the program. The statement can be read as 'for each word in 'fd', write a new line in 'output' which consists of the word itself, then a comma, and then the word's frequency (the '*str(fd[word])*' part), and then start a new line. The program then does this for all the words in 'fd' until they are all in the output file. The reason for the comma and new line is that a CSV file is defined as just a text file with all the values separated by commas, and each new row is signaled by creating a new line. When a CSV file is opened by a spreadsheet program (for example), it thus knows to put each value separated by a comma into a new cell of the spreadsheet and to put each value following a new line into a new row of the spreadsheet.

We therefore have in the output file a list of all the lemmas in the input text and how often they appear. This is formed by counting a lemmatized list of all the words in the input text (minus punctuation, new line characters and stopwords, and ignoring case).

MAKESTRING.PY

```
# MakeString.py
# Author: Marc Alexander (marc@seldomend.com)
# Version: 1.0
# Takes a sorted list of HT categories and converts to a single
string
txt = open('hofstadterheadlist.txt').read()
output = open('hofstadterhtstring.txt', 'w')
txt2 = txt.replace('\'', '')
txt3 = txt2.replace('\n', ' ')
txt4 = txt3.replace('\r', ' ')
print >>output, txt4
output.close()
```

MAKESTRING.PY: ANNOTATED

```
# MakeString.py
# Author: Marc Alexander (marc@seldomend.com)
# Version: 1.0
# Takes a sorted list of HT categories and converts to a single
string
txt = open('hofstadterheadlist.txt').read()
output = open('hofstadterhtstring.txt', 'w')
```

As with WordList.py above, these lines open a file and create a file for output. The input file has to be a list.

```
txt2 = txt.replace('\'', '')
txt3 = txt2.replace('\n', ' ')
txt4 = txt3.replace('\r', ' ')
```

These lines replace all the quote marks and two different types of carriage returns (roughly equivalent to new line and new paragraph) in order to turn the list input into a string.

```
print >>output, txt4
output.close()
```

The content of 'txt4' is output to the file specified above, and the file is closed.

ORDEREDLIST.PY

```
# OrderedList.py
# Author: Marc Alexander (marc@seldomend.com)
# Version: 1.2
# Based on WordList.py v.1.6
import nltk
import re
import string
raw = open('/sautoy.txt').read()
output = open('/sautoyordlist.txt', 'w')
txt = raw.lower()
for punct in string.punctuation:
    txt = txt.replace(punct, ' ')
re.sub(r'\r+', ' ',txt)
stopwords = nltk.corpus.stopwords.words('english')
words = nltk.word_tokenize(txt)
filtered = [f for f in words if f.lower() not in stopwords]
wnl = nltk.WordNetLemmatizer()
lemmas = [wnl.lemmatize(t) for t in filtered]
i = 1
for item in lemmas:
    output.write(str(i) + ',' + item + '\n')
    i=i+1
output.close()
```

ORDEREDLIST.PY: ANNOTATED

```
# OrderedList.py
# Author: Marc Alexander (marc@seldomend.com)
# Version: 1.2
# Based on WordList.py v.1.6
import nltk
import re
import string
raw = open('/sautoy.txt').read()
output = open('/sautoyordlist.txt', 'w')
txt = raw.lower()
for punct in string.punctuation:
    txt = txt.replace(punct, ' ')
re.sub(r'\r+', ' ',txt)
stopwords = nltk.corpus.stopwords.words('english')
words = nltk.word_tokenize(txt)
filtered = [f for f in words if f.lower() not in stopwords]
wnl = nltk.WordNetLemmatizer()
lemmas = [wnl.lemmatize(t) for t in filtered]
```

Until this point, the program is identical to WordList.py.

```
i = 1
```

This defines a new variable, a shortcut name, and gives it the value '1'.

```
for item in lemmas:
    output.write(str(i) + ',' + item + '\n')
    i=i+1
```

For each list item in the list of lemmas, the program does the following:

1. Writes in the output file the value of 'i'.
2. Adds a comma after 'i'.
3. Writes the lemma.
4. Adds a new line character.
5. Adds 1 to the value of 'i'. This way the second line has '2' for the value of 'i', and the third '3' etc.

```
output.close()
```

The output file is closed.

LINE_SPLIT.PY

```
# LineSplit.py
# Author 1: 'sli' on Stack Overflow, http://goo.gl/NLfAU
# Author 2: Marc Alexander (marc@seldomend.com)
# Based on advice at http://goo.gl/BIYxP
# Version: 1.1
splitLen = 30          # lines per file
outputBase = 'sautoy'
input = open('/sautoy.txt', 'r').read().split('\r')
at = 1
for lines in range(0, len(input), splitLen):
    outputData = input[lines:lines+splitLen]
    output = open(outputBase + str(at) + '.txt', 'w')
    output.write('\n'.join(outputData))
    output.close()
    at += 1
```

LINEPLIT.PY: ANNOTATED

```
# LineSplit.py
# Author 1: 'sli' on Stack Overflow, http://goo.gl/NLFAU
# Author 2: Marc Alexander (marc@seldomend.com)
# Based on advice at http://goo.gl/BIYxP
```

Unlike the programs above, this program is heavily based on a Python program given by user 'sli' in response to a question asked by another user on Stack Overflow at the above address in 2009. This user is therefore named as the first author of the program.

```
# Version: 1.1
splitLen = 30          # Number of lines per file
```

These are the number of lines which each output file should have.

```
outputBase = 'sautoy'
```

This is the base name for the output files, used below.

```
input = open('/sautoy.txt', 'r').read().split('\r')
```

This opens the input text.

```
at = 1
```

This is a number counter for each act of splitting, and is used below in naming the program.

```
for lines in range(0, len(input), splitLen):
```

This tells the program to run on all lines between line 0 and the final line. The range() function tells the program to repeat itself over the whole sequence of chunks of the size of splitLen.

```
    outputData = input[lines:lines+splitLen]
```

The program will then take a chunk from the input file of the size of splitLen. Because of the range() function above, this will loop until the end of the file, and so the chunk will increment each time. (That is, the first chunk will be from lines 0-30, the next 31-60, and so on.)

```
    output = open(outputBase + str(at) + '.txt', 'w')
```

This line creates an output file, consisting of outputBase, plus the number of the chunk, plus the file extension 'txt'. So in this example, it would give 'sautoy1.txt'.

```
    output.write('\n'.join(outputData))
```

The program then writes the line chunk to that output file.

```
    output.close()
```

The output file is closed.

```
    at += 1
```

Finally, the at value is increased, ready to start the next chunk.

CFDCORPUSSEARCH.PY

```
# CFDCorpusSearch.py
# Author: Marc Alexander (marc@seldomend.com)
# Version: 1.2
import nltk
from nltk.corpus import PlaintextCorpusReader
corpus_root = '/randallcorp'
wp = PlaintextCorpusReader(corpus_root, '.*')
cfd = nltk.ConditionalFreqDist(
    (target, fileid)
    for fileid in wp.fileids()
    for w in wp.words(fileid)
    for target in ['analogy']
    if w.startswith(target))
cfd.tabulate()
```

CFDCORPUSSEARCH.PY: ANNOTATED

```
# CFDCorpusSearch.py
# Author: Marc Alexander (marc@seldomend.com)
# Version: 1.2
import nltk
from nltk.corpus import PlaintextCorpusReader
```

This opens the NLTK simple plain-text corpus reader.

```
corpus_root = '/randallcorp'
```

This tells the corpus reader where to find the folder of plain text files it should treat as a corpus.

```
wp = PlaintextCorpusReader(corpus_root, '.*')
```

And this tells the corpus reader which files from that folder to use - here, all files are selected (indicated by the .*).

```
cfd = nltk.ConditionalFreqDist(
    (target, fileid)
    for fileid in wp.fileids()
    for w in wp.words(fileid)
    for target in ['analogy']
    if w.startswith(target))
```

Here, a conditional frequency distribution (called cfd) is created, which will count the number of times a variable called target will appear in each file, and also return the fileid (that is, the file name). Then fileid is defined as the file name, target is defined as the word 'analogy', and the program is told to count if any word starts with that target.

```
cfd.tabulate()
```

Finally, the table of the conditional frequency distribution is created.

SQL: LOGLIKELIHOOD DATASHEET

```
SELECT Base.Cmain, First(Base.Heading) AS FirstOfHeading,
Sum(Greene.Occur) AS OcG, 1436794*([OcG]+[OcWIKI])/
(1436794+140054734) AS EFG, 140054734*([OcG]+[OcWIKI])/
(1436794+140054734) AS EFGRef, IIf(IsNull([OcG]),
0,2*([OcG]*Log([OcG]/[EFG]))+([OcWIKI]*Log([OcWIKI]/[EFGRef])))
AS LLG, Sum(Kaku.Occur) AS OcK, 918863*([OcK]+[OcWIKI])/
(918863+140054734) AS EFK, 140054734*([OcK]+[OcWIKI])/
(918863+140054734) AS EFKRef, IIf(IsNull([OcK]),
0,2*([OcK]*Log([OcK]/[EFK]))+([OcWIKI]*Log([OcWIKI]/[EFKRef])))
AS LLK, Sum(Randall.Occur) AS OcR, 1104805*([OcR]+[OcWIKI])/
(1104805+140054734) AS EFR, 140054734*([OcR]+[OcWIKI])/
(1104805+140054734) AS EFRRef, IIf(IsNull([OcR]),
0,2*([OcR]*Log([OcR]/[EFR]))+([OcWIKI]*Log([OcWIKI]/[EFRRef])))
AS LLR, Sum(Sautoy.Occur) AS OcS, 791699*([OcS]+[OcWIKI])/
(791699+140054734) AS EFS, 140054734*([OcS]+[OcWIKI])/
(791699+140054734) AS EFSRef, IIf(IsNull([OcS]),
0,2*([OcS]*Log([OcS]/[EFS]))+([OcWIKI]*Log([OcWIKI]/[EFSRef])))
AS LLS, Sum(Alexander.Occur) AS OcA, 154613*([OcA]+[OcWIKI])/
(154613+140054734) AS EFA, 140054734*([OcA]+[OcWIKI])/
(154613+140054734) AS EFARef, IIf(IsNull([OcA]),
0,2*([OcA]*Log([OcA]/[EFA]))+([OcWIKI]*Log([OcWIKI]/[EFARef])))
AS LLA, Sum(Coppel.Occur) AS OcC, 1049733*([OcC]+[OcWIKI])/
(1049733+140054734) AS EFC, 140054734*([OcC]+[OcWIKI])/
(1049733+140054734) AS EFCRef, IIf(IsNull([OcC]),
0,2*([OcC]*Log([OcC]/[EFC]))+([OcWIKI]*Log([OcWIKI]/[EFCRef])))
AS LLC, Sum(Hofstadter.Occur) AS OcH, 2091256*([OcH]+[OcWIKI])/
(2091256+140054734) AS EFH, 140054734*([OcH]+[OcWIKI])/
(2091256+140054734) AS EFHRef, IIf(IsNull([OcH]),
0,2*([OcH]*Log([OcH]/[EFH]))+([OcWIKI]*Log([OcWIKI]/[EFHRef])))
AS LLH, Sum(REFWiki.Occur) AS OcWIKI
FROM (((((((Base LEFT JOIN Greene ON Base.WordNonOE = Greene.Lemma)
LEFT JOIN Hofstadter ON Base.WordNonOE = Hofstadter.Lemma) LEFT
JOIN Kaku ON Base.WordNonOE = Kaku.Lemma) LEFT JOIN Randall ON
Base.WordNonOE = Randall.Lemma) LEFT JOIN Sautoy ON Base.WordNonOE =
Sautoy.Lemma) LEFT JOIN REFWiki ON Base.WordNonOE =
REFWiki.Lemma) LEFT JOIN Alexander ON Base.WordNonOE =
Alexander.Lemma) LEFT JOIN Coppel ON Base.WordNonOE = Coppel.Lemma
GROUP BY Base.Cmain
ORDER BY Base.Cmain;
```

SQL: LOGLIKELIHOOD DATASHEET: ANNOTATED

The query below begins with a SELECT statement. This specifies the data which will be returned by the query, and takes up the majority of this query.

```
SELECT Base.Cmain, First(Base.Heading) AS FirstOfHeading,
```

It firstly tells the database to return the *HT* category number from the query Base (which itself selects the modern part of the full *HT* database) and the first major heading for that category (giving it 'FirstOfHeading' as a short name).

```
Sum(Greene.Occur) AS OcG,  
1436794*([OcG]+[OcWIKI])/(1436794+140054734) AS EFG,  
140054734*([OcG]+[OcWIKI])/(1436794+140054734) AS EFGRef,  
IIf(IsNull([OcG]),0,2*([OcG]*Log([OcG]/[EFG]))+  
([OcWIKI]*Log([OcWIKI]/[EFGRef]))) AS LLG,
```

This is the main query statement, which repeats for each text. It has the following components, separated by commas:

OcG: The total number of occurrences in the Greene table (which contains the lemmatized list of lexical items in that particular text, *FC*) of lemmas which are in each category of the *HT*. (Note that a later statement, following the SELECT list, specifies that this is per *HT* category.)

EFG: The expected number of activations of that *HT* category in the *FC* text, made up of the total activation size of *FC* (1,436,794 semantic domains) multiplied by the number of occurrences of that domain in both *FC* and the reference corpus, and divided by the sum of the total activations of semantic domains in both *FC* and the reference corpus. (These numbers are here manually listed for simplification of the query.)

EFGRef: The expected number of occurrences of that category in the reference corpus with reference to the *FC* text. Calculated in a similar way to EFG, but replacing the initial figure with the total number of semantic domain activations in the reference corpus (140,054,734).

LLG: The log-likelihood ratio figure for the statistical significance with reference to the likelihood of the EFG value being non-random in comparison to the reference corpus.

```
Sum(Kaku.Occur) AS OcK, 918863*([OcK]+[OcWIKI])/(918863+140054734)  
AS EFK, 140054734*([OcK]+[OcWIKI])/(918863+140054734) AS EFKRef,  
IIf(IsNull([OcK]),0,2*([OcK]*Log([OcK]/[EFK]))+  
([OcWIKI]*Log([OcWIKI]/[EFKRef]))) AS LLK, Sum(Randall.Occur) AS  
OcR, 1104805*([OcR]+[OcWIKI])/(1104805+140054734) AS EFR,  
140054734*([OcR]+[OcWIKI])/(1104805+140054734) AS EFRRef,  
IIf(IsNull([OcR]),0,2*([OcR]*Log([OcR]/[EFR]))+  
([OcWIKI]*Log([OcWIKI]/[EFRRef]))) AS LLR, Sum(Sautoy.Occur) AS  
OcS, 791699*([OcS]+[OcWIKI])/(791699+140054734) AS EFS,  
140054734*([OcS]+[OcWIKI])/(791699+140054734) AS EFSRef,  
IIf(IsNull([OcS]),0,2*([OcS]*Log([OcS]/[EFS]))+  
([OcWIKI]*Log([OcWIKI]/[EFSRef]))) AS LLS, Sum(Alexander.Occur) AS  
OcA, 154613*([OcA]+[OcWIKI])/(154613+140054734) AS EFA,  
140054734*([OcA]+[OcWIKI])/(154613+140054734) AS EFARef,  
IIf(IsNull([OcA]),0,2*([OcA]*Log([OcA]/[EFA]))+  
([OcWIKI]*Log([OcWIKI]/[EFARef]))) AS LLA, Sum(Coppel.Occur) AS  
OcC, 1049733*([OcC]+[OcWIKI])/(1049733+140054734) AS EFC,  
140054734*([OcC]+[OcWIKI])/(1049733+140054734) AS EFCRef,  
IIf(IsNull([OcC]),0,2*([OcC]*Log([OcC]/[EFC]))+  
([OcWIKI]*Log([OcWIKI]/[EFCRef]))) AS LLC, Sum(Hofstadter.Occur)  
AS OcH, 2091256*([OcH]+[OcWIKI])/(2091256+140054734) AS EFH,  
140054734*([OcH]+[OcWIKI])/(2091256+140054734) AS EFHRef,
```

```
IIf(IsNull([Och]),0,2*((([Och]*Log([Och]/[EFH]))+
([OcwIKI]*Log([OcwIKI]/[EFHRef])))) AS LLH,
```

This repeats for each text, the final initial of each item distinguishing the text (Kaku (*PW*), Randall (*WP*), Sautoy (*MP*), Alexander (*C-AC*), Coppel (*C-NT*) and Hofstadter (*C-GEB*)).

```
Sum(REFWiki.Occur) AS OcwIKI
```

This totals the number of activations of each domain in the reference corpus. It is also at this point that the SELECT statement ends and a FROM mapping begins:

```
FROM (((((((Base LEFT JOIN Greene ON Base.WordNonOE = Greene.Lemma)
```

Here begins the list of mappings between each of the data tables involved (one for each text, one for the reference corpus, and one for the *HT*). Each table is somewhat like a large spreadsheet, with rows and columns. This statement establishes a link between the *HT* word field (Base.WordNonOE) and the 'lemma' field of the Greene table (Greene.lemma). It states that the database, when going through the words in the *HT*, should look for matching words in the 'lemma' column of the Greene data. The LEFT part of the LEFT JOIN statement says it only goes one way – the database should go through the *HT* and look at the other tables for each entry in the words of the *HT* table, but not the other way around.

```
LEFT JOIN Hofstadter ON Base.WordNonOE = Hofstadter.Lemma) LEFT
JOIN Kaku ON Base.WordNonOE = Kaku.Lemma) LEFT JOIN Randall ON
Base.WordNonOE = Randall.Lemma) LEFT JOIN Sautoy ON Base.WordNonOE =
Sautoy.Lemma) LEFT JOIN REFWiki ON Base.WordNonOE =
REFWiki.Lemma) LEFT JOIN Alexander ON Base.WordNonOE =
Alexander.Lemma) LEFT JOIN Coppel ON Base.WordNonOE = Coppel.Lemma
```

The same is done for the other tables.

```
GROUP BY Base.Cmain
```

The data is grouped by the category field of the *HT*, meaning that if there are multiple activations (for example, should there be dozens of activations of the *City* domain in each table, the above Sum statements in the SELECT part of the query will operate per *HT* domain, adding together all the occurrences of the *City* domain in each text and giving only a single sum in the returned data, rather than lists of each one separately).

```
ORDER BY Base.Cmain;
```

Finally, the returned data is sorted by the *HT* category number.

SQL: OCCURRENCES OF TOP 100 DOMAINS

```
SELECT OrdAlexander.Ord, OrdAlexander.Lemma, Base.Cmain
FROM Base INNER JOIN OrdAlexander ON Base.WordNonOE =
OrdAlexander.Lemma
GROUP BY OrdAlexander.Ord, OrdAlexander.Lemma, Base.Cmain
HAVING (((Base.Cmain)='02.08.03.06.02')) OR
(((Base.Cmain)='03.08.08.11.01')) OR
(((Base.Cmain)='03.07.02.05.12.02')) OR
(((Base.Cmain)='02.01.13.08.02')) OR
(((Base.Cmain)='03.11.04.13.12.01')) OR
(((Base.Cmain)='03.11.03.04.03.02')) OR
(((Base.Cmain)='03.08.09.01')) OR (((Base.Cmain)='02.08.03.06.01'))
OR (((Base.Cmain)='02.01.10.02.03.03')) OR
(((Base.Cmain)='03.08.03.03.01.01')) OR (((Base.Cmain)='03.08.09'))
OR (((Base.Cmain)='03.08.08.23.02')) OR
(((Base.Cmain)='03.11.01.05.02.07.13')) OR
(((Base.Cmain)='02.08.04.04')) OR
(((Base.Cmain)='02.01.07.04.03.02')) OR
(((Base.Cmain)='01.05.08.01')) OR (((Base.Cmain)='03.11.03.01.07'))
OR (((Base.Cmain)='03.06.02.05.02')) OR
(((Base.Cmain)='02.01.15.06.01')) OR
(((Base.Cmain)='01.02.02.04.01.01')) OR
(((Base.Cmain)='03.11.03.03.07.04')) OR
(((Base.Cmain)='02.08.03.03')) OR (((Base.Cmain)='01.03.07.08.02'))
OR (((Base.Cmain)='02.01.11')) OR
(((Base.Cmain)='03.11.01.05.02.01')) OR
(((Base.Cmain)='03.04.13.12.06.03')) OR
(((Base.Cmain)='03.08.07.06')) OR
(((Base.Cmain)='01.02.05.20.05.01')) OR (((Base.Cmain)='02.05.02'))
OR (((Base.Cmain)='01.05.05.20.07')) OR
(((Base.Cmain)='02.01.12.02.01')) OR
(((Base.Cmain)='01.06.04.15.01.01')) OR
(((Base.Cmain)='01.05.06.08.05.03')) OR (((Base.Cmain)='02.08.03'))
OR (((Base.Cmain)='03.11.03.03.07.04.06')) OR
(((Base.Cmain)='03.11.03.03.08.02.01')) OR
(((Base.Cmain)='03.07.00.20.01')) OR
(((Base.Cmain)='01.05.07.04.06.01')) OR (((Base.Cmain)='01.05.08'))
OR (((Base.Cmain)='02.01.15.01')) OR (((Base.Cmain)='02.05.04')) OR
(((Base.Cmain)='03.04.06.16')) OR (((Base.Cmain)='01.05.06.04.01'))
OR (((Base.Cmain)='03.11.03.03.01.02')) OR
(((Base.Cmain)='03.11.03.03.05.08')) OR (((Base.Cmain)='02.01.07'))
OR (((Base.Cmain)='01.04.05.01.02')) OR
(((Base.Cmain)='02.01.13.08.06')) OR (((Base.Cmain)='03.06.06.02'))
OR (((Base.Cmain)='02.05.04.01.01')) OR
(((Base.Cmain)='01.03.07.03')) OR (((Base.Cmain)='02.01.04')) OR
(((Base.Cmain)='02.01.11.02')) OR
(((Base.Cmain)='03.11.03.03.01.01')) OR
(((Base.Cmain)='03.08.07.05')) OR (((Base.Cmain)='03.07.00.18.01'))
OR (((Base.Cmain)='03.06.03.01')) OR (((Base.Cmain)='02.01.10.03'))
OR (((Base.Cmain)='02.01.13.04')) OR (((Base.Cmain)='02.02.03')) OR
(((Base.Cmain)='02.01.08')) OR
```

((Base.Cmain)='02.01.12.08.06.02.01') OR
((Base.Cmain)='01.02.02.04.10') OR ((Base.Cmain)='02.02.20.09'))
OR ((Base.Cmain)='01.05.05.21.04.01.02') OR
((Base.Cmain)='02.01.05.18') OR ((Base.Cmain)='03.07.03.11') OR
((Base.Cmain)='01.06.07.05') OR
((Base.Cmain)='03.11.03.04.03.04') OR
((Base.Cmain)='03.11.03.04.02.01') OR
((Base.Cmain)='02.08.02.06') OR ((Base.Cmain)='02.01.12') OR
((Base.Cmain)='02.01.12.08.06') OR ((Base.Cmain)='01.06.04.12'))
OR ((Base.Cmain)='02.01.03.06') OR ((Base.Cmain)='03.03.18.01'))
OR ((Base.Cmain)='01.02.09.15.03.01') OR
((Base.Cmain)='03.01.04.03') OR ((Base.Cmain)='02.01.05.10')) OR
((Base.Cmain)='03.11.05.06') OR
((Base.Cmain)='02.01.17.04.02.03') OR
((Base.Cmain)='03.08.07.04') OR
((Base.Cmain)='02.01.05.16.02.01') OR
((Base.Cmain)='02.01.12.08.06.06') OR
((Base.Cmain)='01.06.02.02.01') OR
((Base.Cmain)='02.01.11.02.01') OR
((Base.Cmain)='03.11.03.01.09') OR ((Base.Cmain)='02.01.15.05'))
OR ((Base.Cmain)='03.11.03.03.05') OR
((Base.Cmain)='03.11.03.03.08.01') OR
((Base.Cmain)='03.02.07.01.04.03') OR
((Base.Cmain)='02.01.10.02.03') OR
((Base.Cmain)='03.02.07.01.04.02') OR
((Base.Cmain)='03.02.07.01.01') OR
((Base.Cmain)='03.11.03.02.04.03') OR
((Base.Cmain)='03.07.02.04.19') OR ((Base.Cmain)='02.02.20.06'))
OR ((Base.Cmain)='03.08.04') OR ((Base.Cmain)='02.03.28.02') OR
((Base.Cmain)='03.11.03.04.17.02.03'))
ORDER BY Base.Cmain

SQL: OCCURRENCES OF TOP 100 DOMAINS: ANNOTATED

```
SELECT OrdAlexander.Ord, OrdAlexander.Lemma, Base.Cmain
```

This tells the query to give an output which consists of the lemma and where in the text that lemma occurs (Ord) of the data table for that text (in this example, the Alexander C-AC ordered list table), and the *HT* category numbers corresponding to each lemma. There is one query of this type per table.

```
FROM Base INNER JOIN OrdAlexander ON Base.WordNonOE =  
OrdAlexander.Lemma
```

Here the necessary table mapping is established. By contrast with the LogLikelihood query above, this uses an INNER JOIN keyword, which will only return results when there are results in both tables (that is, it will not return *HT* entries with no corresponding lemma, or lemmas with no *HT* entries, such as proper names). The link ties the text lemma to the *HT* word field.

```
GROUP BY OrdAlexander.Ord, OrdAlexander.Lemma, Base.Cmain
```

The data is grouped by the lemma order, then by the lemma itself, then by the category fields of the *HT* belonging to each lemma. As in the LogLikelihood table above, this ensures that multiple activations will be totaled per *HT* domain, and also per lemma.

```
HAVING (((Base.Cmain)='02.08.03.06.02')) OR  
(((Base.Cmain)='03.08.08.11.01')) OR  
(((Base.Cmain)='03.07.02.05.12.02')) OR  
(((Base.Cmain)='02.01.13.08.02')) OR  
(((Base.Cmain)='03.11.04.13.12.01')) OR  
(((Base.Cmain)='03.11.03.04.03.02')) OR  
(((Base.Cmain)='03.08.09.01')) OR (((Base.Cmain)='02.08.03.06.01'))  
OR (((Base.Cmain)='02.01.10.02.03.03')) OR  
(((Base.Cmain)='03.08.03.03.01.01')) OR (((Base.Cmain)='03.08.09'))  
OR (((Base.Cmain)='03.08.08.23.02')) OR  
(((Base.Cmain)='03.11.01.05.02.07.13')) OR  
(((Base.Cmain)='02.08.04.04')) OR  
(((Base.Cmain)='02.01.07.04.03.02')) OR  
(((Base.Cmain)='01.05.08.01')) OR (((Base.Cmain)='03.11.03.01.07'))  
OR (((Base.Cmain)='03.06.02.05.02')) OR  
(((Base.Cmain)='02.01.15.06.01')) OR  
(((Base.Cmain)='01.02.02.04.01.01')) OR  
(((Base.Cmain)='03.11.03.03.07.04')) OR  
(((Base.Cmain)='02.08.03.03')) OR (((Base.Cmain)='01.03.07.08.02'))  
OR (((Base.Cmain)='02.01.11')) OR  
(((Base.Cmain)='03.11.01.05.02.01')) OR  
(((Base.Cmain)='03.04.13.12.06.03')) OR  
(((Base.Cmain)='03.08.07.06')) OR  
(((Base.Cmain)='01.02.05.20.05.01')) OR (((Base.Cmain)='02.05.02'))  
OR (((Base.Cmain)='01.05.05.20.07')) OR  
(((Base.Cmain)='02.01.12.02.01')) OR  
(((Base.Cmain)='01.06.04.15.01.01')) OR  
(((Base.Cmain)='01.05.06.08.05.03')) OR (((Base.Cmain)='02.08.03'))  
OR (((Base.Cmain)='03.11.03.03.07.04.06')) OR  
(((Base.Cmain)='03.11.03.03.08.02.01')) OR  
(((Base.Cmain)='03.07.00.20.01')) OR  
(((Base.Cmain)='01.05.07.04.06.01')) OR (((Base.Cmain)='01.05.08'))  
OR (((Base.Cmain)='02.01.15.01')) OR (((Base.Cmain)='02.05.04')) OR
```

```

(((Base.Cmain)='03.04.06.16')) OR (((Base.Cmain)='01.05.06.04.01'))
OR (((Base.Cmain)='03.11.03.03.01.02')) OR
(((Base.Cmain)='03.11.03.03.05.08')) OR (((Base.Cmain)='02.01.07'))
OR (((Base.Cmain)='01.04.05.01.02')) OR
(((Base.Cmain)='02.01.13.08.06')) OR (((Base.Cmain)='03.06.06.02'))
OR (((Base.Cmain)='02.05.04.01.01')) OR
(((Base.Cmain)='01.03.07.03')) OR (((Base.Cmain)='02.01.04')) OR
(((Base.Cmain)='02.01.11.02')) OR
(((Base.Cmain)='03.11.03.03.01.01')) OR
(((Base.Cmain)='03.08.07.05')) OR (((Base.Cmain)='03.07.00.18.01'))
OR (((Base.Cmain)='03.06.03.01')) OR (((Base.Cmain)='02.01.10.03'))
OR (((Base.Cmain)='02.01.13.04')) OR (((Base.Cmain)='02.02.03')) OR
(((Base.Cmain)='02.01.08')) OR
(((Base.Cmain)='02.01.12.08.06.02.01')) OR
(((Base.Cmain)='01.02.02.04.10')) OR (((Base.Cmain)='02.02.20.09'))
OR (((Base.Cmain)='01.05.05.21.04.01.02')) OR
(((Base.Cmain)='02.01.05.18')) OR (((Base.Cmain)='03.07.03.11')) OR
(((Base.Cmain)='01.06.07.05')) OR
(((Base.Cmain)='03.11.03.04.03.04')) OR
(((Base.Cmain)='03.11.03.04.02.01')) OR
(((Base.Cmain)='02.08.02.06')) OR (((Base.Cmain)='02.01.12')) OR
(((Base.Cmain)='02.01.12.08.06')) OR (((Base.Cmain)='01.06.04.12'))
OR (((Base.Cmain)='02.01.03.06')) OR (((Base.Cmain)='03.03.18.01'))
OR (((Base.Cmain)='01.02.09.15.03.01')) OR
(((Base.Cmain)='03.01.04.03')) OR (((Base.Cmain)='02.01.05.10')) OR
(((Base.Cmain)='03.11.05.06')) OR
(((Base.Cmain)='02.01.17.04.02.03')) OR
(((Base.Cmain)='03.08.07.04')) OR
(((Base.Cmain)='02.01.05.16.02.01')) OR
(((Base.Cmain)='02.01.12.08.06.06')) OR
(((Base.Cmain)='01.06.02.02.01')) OR
(((Base.Cmain)='02.01.11.02.01')) OR
(((Base.Cmain)='03.11.03.01.09')) OR (((Base.Cmain)='02.01.15.05'))
OR (((Base.Cmain)='03.11.03.03.05')) OR
(((Base.Cmain)='03.11.03.03.08.01')) OR
(((Base.Cmain)='03.02.07.01.04.03')) OR
(((Base.Cmain)='02.01.10.02.03')) OR
(((Base.Cmain)='03.02.07.01.04.02')) OR
(((Base.Cmain)='03.02.07.01.01')) OR
(((Base.Cmain)='03.11.03.02.04.03')) OR
(((Base.Cmain)='03.07.02.04.19')) OR (((Base.Cmain)='02.02.20.06'))
OR (((Base.Cmain)='03.08.04')) OR (((Base.Cmain)='02.03.28.02')) OR
(((Base.Cmain)='03.11.03.04.17.02.03'))

```

This list, automatically generated in Python, simply means 'return results where the *HT* category number is one of these'. There are simpler ways of doing this (using an IN statement), but this is the least ambiguous for presentation purposes. The *HT* categories are the most frequent activations, found using the results of the LogLikelihood datasheet. These are different for each text, for obvious reasons.

ORDER BY Base.Cmain

Finally, the table is sorted by *HT* category number.

APPENDIX II.1: VITAL RELATIONS IN CONCEPTUAL INTEGRATION THEORY

The CIT Vital Relations are listed here in the order they are detailed in Fauconnier and Turner (2002:93ff).

Vital Relation	Description from Fauconnier and Turner 2002	Page Numbers
Change	connecting one element to another and suites of elements to other suites [...] Conceptually, a sapling and the tree it grows into set up two mental spaces connected by Change. Age changes a person, translation changes a text, ‘Americanizing’ something foreign changes it into something suitable for Americans. Mental spaces are dynamic, so change can be located within an individual mental space’	93-94
Identity	blending is a powerful and supple instrument for creating and disintegrating identity [...] identity connectors always involve interesting differences: For example, one monk is a few days older than the other, just as ‘Mary’ as a baby and ‘Mary’ as a fifty-year-old CEO and grandmother are linked by an identity connector across temporally separate spaces [...] identity across spaces is a stipulated connection.’	95-96
Time	‘related to memory, change, continuity, simultaneity, and nonsimultaneity, as well as to our understanding of causation’	96
Space	Many blend networks ‘have inputs separated in physical space but a blend with a single physical space. Blends very frequently compress over Space.’	96

Vital Relation	Description from Fauconnier and Turner 2002	Page Numbers
Cause-Effect	<p>‘It is not trivial to bring cause and effect together. They have to be brought together in one mental space, in the right way, while being kept distinct in other spaces [...]In perception, at the level of consciousness, we usually apprehend only the blend of cause and effect. We cannot fail to perform this blend, and we cannot see beyond it. Consequently, this blend seems to us to be the most bedrock reality [...] It is not enough to see one thing as caused by the other; rather, we need the two proper mental spaces.’</p>	75-87, 96
Part-Whole	<p>We point to a picture of a face and say ‘That’s Jane Doe,’ not ‘That’s the face of Jane Doe’. We have constructed a network mapping the individual to the picture of what seems to us her most salient part, her face. In the blend, the face is projected from one input and the whole person is projected from the other. In the blend, face and person are fused: The face is the personal identity. The part-whole connection between the face and person in the input spaces becomes uniqueness in the blend.’</p>	97

Vital Relation	Description from Fauconnier and Turner 2002	Page Numbers
Representation	<p>One input can have a representation of the other – as in a sketch of a person or a picture of a baby [...] In the blend, the representation link between the thing represented and the thing representing it is typically compressed into the uniqueness [...] When we look at a picture of the face of Jane and say ‘this is Jane,’ we are actually using a complicated network with multiple blending and compression.’ Fauconnier and Turner explain this as a combination of a representation of Jane’s face being a representation of Jane herself (Part-Whole) and a representation of Jane being compressed into uniqueness with Jane.</p>	97-8
Role	<p>‘Lincoln was <i>president</i>, Elizabeth is <i>queen</i>, and the president is the <i>head of state</i>: Roles have values. Lincoln, in 1863, is a value for president; Elizabeth, today, is a value for queen; and president, in the United States, is a value for head of state. Within mental spaces, and across mental spaces, an element can be linked, and a role, to another element that counts as its value. Elements are roles or values not in some absolute sense but only relative to other elements.’</p>	98
Analogy	<p>Analogy depends upon Role-Value compression [...] Stanford is a West Coast ‘analogue’ to Harvard. Both blending networks have the same American university frame with the role <i>prestigious private American research university</i>. One network has the value Stanford in the other input; the other network has a different value, Harvard, in the other input, The blends in the two networks are linked by Analogy because of the identity of those input roles.’</p>	98-99

Vital Relation	Description from Fauconnier and Turner 2002	Page Numbers
Disanalogy	<p>‘Disanalogy is grounded on Analogy. We are not disposed to think of a brick and the Atlantic Ocean as disanalogous, but we are disposed to think of the Atlantic Ocean and the Pacific Ocean as disanalogous. Disanalogy is coupled to Analogy [...] Disanalogy is often compressed into Change.’ Here Fauconnier and Turner’s implied definition is that two entities are joined with similar features but different roles (the Atlantic and Pacific are very similar but do different ‘jobs’ in different places) whereas Analogy has two entities with similar roles <i>and</i> similar features.</p>	99
Property	<p>A blue cup has the property <i>blue</i>. A saint has the property <i>holy</i>. A murderer has the property <i>guilty</i> [...]</p> <p>The most obvious status of a property is as an inner-space vital relation: In the space of the blue cup, the cup is intrinsically blue. Blending often compresses an outer-space vital relation of some sort into an inner-space relation of property in the blend. For example, outer-space Cause-Effect links can be compressed into Property relations in the blend. A warm coat is something that causes you to be warm; it is not something that is warm itself, but in the blend, it is has [<i>sic</i>] the Property <i>warm</i>.’</p>	99-100
Similarity	<p>‘...an inner-space Vital Relation linking elements with shared properties [...] More diffuse outer-space links in mental space networks can be compressed into Similarity in the human-scale blend. For example, outer-space Analogy can be compressed into direct Similarity in the blend.’</p>	100

Vital Relation	Description from Fauconnier and Turner 2002	Page Numbers
Category	<p>‘The most obvious status of a category is as an inner-space vital relation [...] Blending can compress outer-space Vital Relations such as Analogy into Category in the blend. For example, what starts out as an outer-space analogy between a biological virus and an unwanted destructive computer program that invisibly comes to reside on your computer is compressed into a Category relation in the blend: In the blend, the computer program is a <i>virus</i>.’</p>	100
Intentionality	<p>Intentionality covers a group of vital relations having to do with hope, desire, want, fear, belief, memory and other mental attitudes and dispositions directed at content: We fear it will rain, hope we will get home, believe we are in California, remember that we were in France. We interpret each other on the basis of the view that people’s actions and reactions are intentional in this technical sense. Intentionality is crucial because everything we do and think and feel is based on the relations it covers. It makes a difference to us whether the glass breaks accidentally or because we meant to break it.’</p>	100-101
Uniqueness	<p>Uniqueness obtains automatically for elements in the blend [<i>sic</i>], and we take it for granted. The importance of Uniqueness in a technical sense is that many vital relations compress into Uniqueness in the blend.’ This is all Fauconnier and Turner say about Uniqueness.</p>	101

APPENDIX II.2: *PARALLEL WORLDS*

a: Full Text

The Holographic Universe

Last, there is a rather mysterious prediction of M-theory that is still not understood but may have deep physical and philosophical consequences. This result forces us to ask the question: is the universe a hologram? Is there a 'shadow universe' in which our bodies exist in a compressed two-dimensional form? This also raises another, equally disturbing question: is the universe a computer program? Can the universe be placed on a CD, to be played at our leisure?

Holograms are now found on credit cards, in children's museums, and in amusement parks. They are remarkable because they can capture a complete three-dimensional image on a two-dimensional surface. Normally, if you glance at a photograph and then move your head, the image on the photograph does not change. But a hologram is different. When you glance at a holographic picture and then move your head, you find the picture changing, as if you were looking at the image through a window or a keyhole. (Holograms may eventually lead to three-dimensional TV and movies. In the future, perhaps we will relax in our living room and glance at a wall screen that gives us the complete three-dimensional image of distant locations, as if the TV wall screen were actually a window peering out over a new landscape. Furthermore, if the wall screen were shaped like a large cylinder with our living room placed in the center, it would appear as if we were transported to a new world. Everywhere we looked, we would see the three-dimensional image of a new reality, indistinguishable from the real thing.)

The essence of the hologram is that the two-dimensional surface of the hologram encodes all the information necessary to reproduce a three-

dimensional image. (Holograms are made in the laboratory by shining laser light onto a sensitive photographic plate and allowing the light to interfere with laser light from the original source. The interference of the two light sources creates an interference pattern that ‘freezes’ the image onto the two-dimensional plate.)

Some cosmologists have conjectured that this may also apply to the universe itself – that perhaps we live in a hologram. The origin of this strange speculation arises from black hole physics. Bekenstein and Hawking conjecture that the total amount of information contained in a black hole is proportional to the surface area of its event horizon (which is a sphere). This is a strange result, because usually the information stored in an object is proportional to its volume. For example, the amount of information stored in a book is proportional to its size, not to the surface area of its cover. We know this instinctively, when we say that we cannot judge a book by its cover. But this intuition fails for black holes: we can completely judge a black hole by its cover.

We may dismiss this curious hypothesis because black holes are strange oddities in themselves, where normal intuition breaks down. However, this result also applies to M-theory, which may give us the best description of the entire universe. In 1997, Juan Maldacena, at the Institute for Advanced Study at Princeton, created quite a sensation when he showed that string theory leads to a new type of holographic universe.

He started with a five-dimensional ‘anti-de Sitter universe’ which often appears in string theory and supergravity theory. A de Sitter universe is one with a positive cosmological constant that creates an accelerating universe. (We recall that our universe is currently best represented as a de Sitter universe, with a cosmological constant pushing the galaxies away at faster and faster velocities. An anti-de Sitter universe has a negative cosmological constant and hence can implode.) Maldacena showed that there is a duality between this five-dimensional universe and its ‘boundary,’ which is a four-dimensional universe.

Strangely enough, any beings living in this five-dimensional space would be mathematically equivalent to beings living in this four-dimensional space. There is no way to tell them apart.

By crude analogy, think of fish swimming inside a goldfish bowl. These fish think that their fish bowl corresponds to reality. Now imagine a two-dimensional holographic image of these fish that is projected onto the surface of the fish bowl. This image contains an exact replica of the original fish, except they are flattened. Any movement the fish make in the fish bowl is mirrored by the flat image on the surface of the fish bowl. Both the fish swimming in the bowl and the flattened fish living on the surface of the bowl think that they are the real fish, that the other is an illusion. Both fish are alive and act as if they are the true fish. Which description is correct? Actually, both are, since they are mathematically equivalent and indistinguishable.

What excited string theorists is the fact that five-dimensional anti-de Sitter space is relatively easy to calculate with, while four-dimensional field theories are notoriously difficult to handle. (Even today, after decades of hard work, our most powerful computers cannot solve the four-dimensional quark model and derive the masses of the proton and neutron. The equations for the quarks themselves are fairly well understood, but solving them in four dimensions to obtain the properties of protons and neutrons has proved to be more difficult than previously thought.) One goal is to calculate the masses and properties of the proton and neutron, using this strange duality.

This holographic duality may also have practical applications, such as solving the information problem in black hole physics. In four dimensions, it is extremely difficult to prove that information isn't lost when we throw objects through a black hole. But such a space is dual to a five-dimensional world, in which information is perhaps never lost. The hope is that problems that are intractable in four dimensions (such as the information problem, calculating the masses of the quark model, and so forth) may eventually be solved in five

dimensions, where the mathematics is simpler. And it is always possible that this analogy is actually a reflection of the real world – that we really exist as holograms.

b: Annotated Text

The Holographic Universe

Last, there is a rather mysterious *prediction* > of **M-theory** that is still not understood but may have deep physical and philosophical consequences. This result forces us to ask the question: is the *universe* a **hologram**? Is there a ‘**shadow universe**’ in which our bodies exist in a compressed two-dimensional form? This also raises another, equally disturbing question: is the *universe* a **computer program**? Can the universe be placed on a CD, to be played at our leisure?

Holograms are now found on credit cards, in children’s museums, and in amusement parks. They are remarkable because they can capture a complete three-dimensional image on a two-dimensional surface. Normally, if you glance at a **photograph** and then move your head, the image on the photograph does not change. But a *hologram* is different. When you glance at a *holographic picture* and then move your head, you find the picture changing, as if you were looking at the *image* through a **window** or a **keyhole**. [Parenthesis omitted.]

The essence of the *hologram* is that the two-dimensional surface of the *hologram* encodes all the information necessary to reproduce a three-dimensional *image*. [Parenthesis omitted.]

[Circumstantial agency omitted] perhaps we live in a *hologram*. The origin of *this strange speculation* arises from **black hole physics**. [Circumstantial agency omitted] the total amount of **information** contained in a *black hole* is proportional to the surface area of its event horizon (which is a sphere). This is *a strange result*, because usually the *information* stored in an **object** is

proportional to its volume. For example, the amount of *information* stored in a **book** is proportional to its size, not to the surface area of its cover. We know *this* instinctively, when we say that we cannot judge a *book* by its cover. But this *intuition* fails for *black holes*: we can completely judge a *black hole* by its *cover*.

We may dismiss *this curious hypothesis* because *black holes* are strange oddities in themselves, where normal *intuition* breaks down. However, *this result* also applies to *M-theory*, which may give us the best description of the entire *universe*. [Circumstantial agency omitted] **string theory** leads to a new type of *holographic universe*.

He started with a five-dimensional ‘**anti-de Sitter universe**’ which often appears in *string theory* and *supergravity theory*. A **de Sitter universe** is one with a positive cosmological constant that creates an accelerating universe.

[Parenthesis omitted.] [Circumstantial agency omitted] there is a *duality* > between *this five-dimensional universe* and its ‘**boundary**,’ which is a four-dimensional universe. Strangely enough, any beings living in *this five-dimensional space* would be mathematically equivalent to beings living in *this four-dimensional space*. There is no way to tell them apart.

By crude *analogy* >, think of **fish** swimming inside a **goldfish bowl**. These *fish* think that their *fish bowl* corresponds to **reality**. Now imagine a **two-dimensional holographic image of these fish** that is projected onto the surface of the *fish bowl*. This *image* contains an exact replica of the original fish, except they are flattened. Any movement the *fish* make in the *fish bowl* is mirrored by the flat *image* on the *surface* of the *fish bowl*. Both the *fish* swimming in the *bowl* and the *flattened fish* living on the *surface* of the *bowl* think that they are the real fish, that the other is an illusion. Both *fish* are alive and act as if they are the true fish. Which description is correct? Actually, both are, since they are mathematically equivalent and indistinguishable.

[Circumstantial agency omitted] *five-dimensional anti-de Sitter space* is relatively easy to calculate with, while *four-dimensional field theories* are notoriously difficult to handle. [Parenthesis omitted.] One goal is to calculate the masses and properties of the **proton** and **neutron**, using this strange *duality*.

This holographic *duality* may also have practical applications, such as solving the **information problem** in *black hole physics*. In *four dimensions*, it is extremely difficult to prove that *information isn't lost when we throw objects through a black hole*. But such a *space* is dual to a *five-dimensional world*, in which information is perhaps never lost. The hope is that **problems** that are intractable in *four dimensions* (such as the *information problem*, calculating the masses of the quark model, and so forth) may eventually be solved in *five dimensions*, where the mathematics is simpler. And it is always possible that this *analogy* is actually a reflection of the *real world* – that we really exist as *holograms*.

c: Input Mental Spaces

Identifier	First enactor	Relevant <i>HT</i> category	Space name
I ₁	universe	O1O110: <i>Universe</i>	O1O110: <i>Universe</i>
I ₂	M-theory	[see §5.ii.1.b in text]	O1O40707: <i>Quantum/string theory</i>
I ₃	hologram	O1O40703: <i>Electromagnetic radiation</i>	O1O40703: <i>Hologram</i>
I ₄	shadow universe	O1O110: <i>Universe</i>	O1O110: <i>Shadow universe</i>
I ₅	computer program	O1O60415: <i>Computing/information technology</i>	O1O60415: <i>Computer program</i>
I ₆	photograph	O3110302: <i>Visual arts</i>	O3110302: <i>Photograph</i>
I ₇	window	O1O50705: <i>Relative position</i>	O1O50705: <i>Window</i>
I ₈	keyhole	O1O50705: <i>Relative position</i>	O1O50705: <i>Keyhole</i>

Identifier	First enactor	Relevant <i>HT</i> category	Space name
I ₉	black hole physics	01011006: <i>Star</i>	01011006: <i>Black hole</i>
I ₁₀	information	030805: <i>Information</i>	030805: <i>Information</i>
I ₁₁	object	030802: <i>Manifestation</i>	030802: <i>Object</i>
I ₁₂	book	03070416: <i>Book (general)</i>	03070416: <i>Book</i>
I ₁₃	anti-de Sitter universe	[see §5.ii.1.b in text]	011010: <i>Anti-de Sitter universe</i>
I ₁₄	de Sitter universe	[see §5.ii.1.b in text]	011010: <i>de Sitter universe</i>
I ₁₅	boundary	01050705: <i>Relative position</i>	01050705: <i>Boundary</i>
I ₁₆	fish	01020616: <i>Fish</i>	01020616: <i>Fish</i>
I ₁₇	goldfish bowl	03101140: <i>Receptacle/container</i>	03101140: <i>Goldfish bowl</i>
I ₁₈	reality	01050102: <i>Reality/real existence/actuality</i>	01050102: <i>Reality</i>
I ₁₉	holographic image of these fish	01040703: <i>Electromagnetic radiation</i>	01040703: <i>Hologram (of fish)</i>
I ₂₀	proton	01040705: <i>Atomic physics</i>	01040705: <i>Proton</i>
I ₂₁	neutron	01040705: <i>Atomic physics</i>	01040705: <i>Neutron</i>
I ₂₂	information problem	01060402: <i>Mathematics</i>	01060402: <i>Information problem</i>
I ₂₃	problems	01060402: <i>Mathematics</i>	01060402: <i>Mathematical problems</i>

d: Semantic Input Space Groups

010110 UNIVERSE

010110: *Anti-de Sitter universe*

010110: *de Sitter universe*

010110: *Shadow universe*

010110: *Universe*

01011006 BLACK HOLE

01011006: *Black hole*

01020616 FISH

01020616: *Fish*

010407 PHYSICS

01040703: *Hologram*

01040703: *Hologram (of fish)*

01040705: *Neutron*

01040705: *Proton*

01040707: *Quantum/string theory*

0105 EXISTENCE, FORM

01050102: *Reality*

01050705: *Boundary*

01050705: *Keyhole*

01050705: *Window*

01060402 MATHEMATICS

01060402: *Information problem*

01060402: *Mathematical problems*

01060415: *Computer program*

03 OBJECTS IN THE WORLD

03070416: *Book*

030802: *Object*

03101140: *Goldfish bowl*

03110302: *Photograph*

030805 INFORMATION

030805: *Information*

e: Grouped Mental Spaces

Identifier	Space name	Semantic grouping	Textual grouping
I ₁	010110: <i>Universe</i>	010110 UNIVERSE	
I ₂	01040707: <i>Quantum/string theory</i>	010407 PHYSICS	
I ₃	01040703: <i>Hologram</i>	010407 PHYSICS	
I ₄	010110: <i>Shadow universe</i>	010110 UNIVERSE	
I ₅	01060415: <i>Computer program</i>	01060402 MATHEMATICS	
I ₆	03110302: <i>Photograph</i>	03 OBJECTS IN THE WORLD	
I ₇	01050705: <i>Aperture</i>	0105 EXISTENCE, FORM	
I ₉	01011006: <i>Black hole</i>	01011006 BLACK HOLE	BLACK HOLE
I ₁₀	030805: <i>Information</i>	030805 INFORMATION	BLACK HOLE
I ₁₁	030802: <i>Object</i>	03 OBJECTS IN THE WORLD	BLACK HOLE

Identifier	Space name	Semantic grouping	Textual grouping
I ₁₂	030704I6: <i>Book</i>	03 OBJECTS IN THE WORLD	BLACK HOLE
I ₁₃	011010: <i>Anti-de Sitter universe</i>	010110 UNIVERSE	ANTI-DE SITTER UNIVERSE
I ₁₄	011010: <i>de Sitter universe</i>	010110 UNIVERSE	ANTI-DE SITTER UNIVERSE
I ₁₅	01050705: <i>Boundary</i>	0105 EXISTENCE, FORM	ANTI-DE SITTER UNIVERSE
I ₁₆	010206I6: <i>Fish</i>	010206I6 FISH	FISH
I ₁₇	03101140: <i>Goldfish bowl</i>	03 OBJECTS IN THE WORLD	FISH
I ₁₈	01050102: <i>Reality</i>	0105 EXISTENCE, FORM	FISH
I ₁₉	01040703: <i>Hologram (of fish)</i>	010407 PHYSICS	FISH
I ₂₀	01040705: <i>Baryon</i>	010407 PHYSICS	
I ₂₂	01060402: <i>Information problem</i>	01060402 MATHEMATICS	MATHEMATICS
I ₂₃	01060402: <i>Mathematical problems</i>	01060402 MATHEMATICS	MATHEMATICS

f: Mental Spaces and Elements

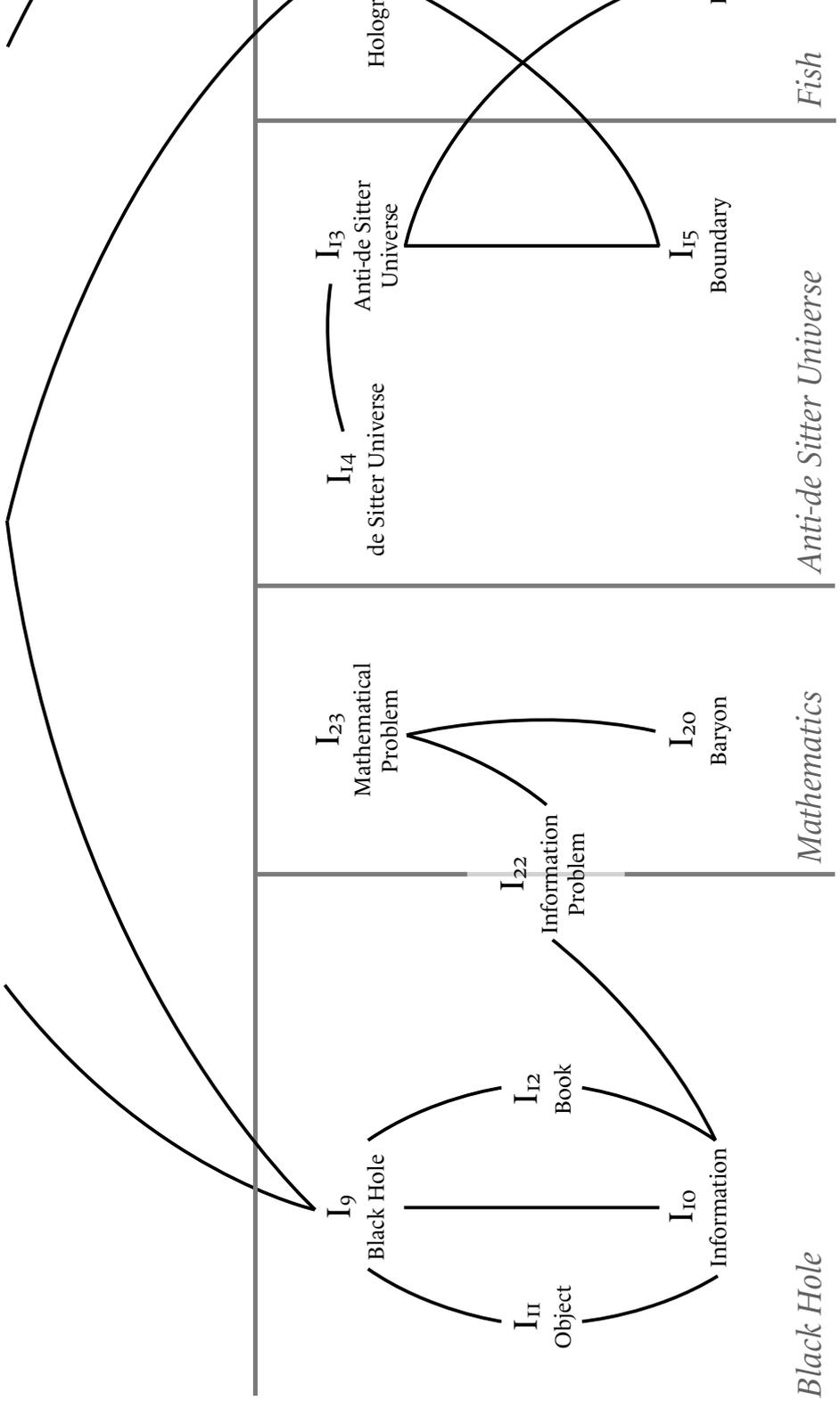
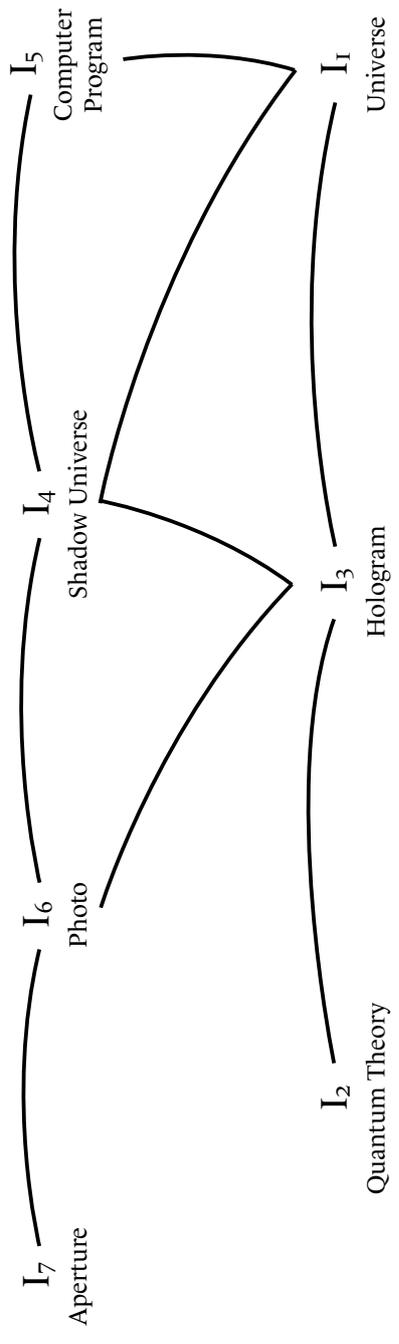
Identifier	Space name	Elements (paraphrased for concision)
I ₁	010110: <i>Universe</i>	Possibly a Hologram
I ₂	01040707: <i>Quantum theory</i>	Requires holographic universe
I ₃	01040703: <i>Hologram</i>	<p>Captures 3d image on 2d surface</p> <p>2d surface encodes all information necessary to reproduce 3d image</p> <p>Movement while observing results in change in observed image</p> <p>Found on/in credit cards, museums, and amusement parks</p>

Identifier	Space name	Elements (paraphrased for concision)
I ₄	010110: <i>Shadow universe</i>	3d Objects exist in compressed 2d form
I ₅	01060415: <i>Computer program</i>	Can be placed on CD
I ₆	03110302: <i>Photograph</i>	3d Objects represented as if 2d Movement while observing results in no change
I ₇	01050705: <i>Aperture</i>	Such as window or keyhole
I ₉	01011006: <i>Black hole</i>	Has spherical event horizon Contains Information proportional to the surface area of this event horizon
I ₁₀	030805: <i>Information</i>	Usually Information stored in an Object is proportional to volume
I ₁₁	030802: <i>Object</i>	[Elements are solely schematic]
I ₁₂	03070416: <i>Book</i>	Information stored in this Object is proportional to size, not cover surface area Cannot be judged by cover
I ₁₃	011010: <i>Anti-de Sitter universe</i>	5d Has negative cosmological constant Collapsing form of Universe Has beings Easy to calculate/analyse
I ₁₄	011010: <i>de Sitter universe</i>	5d Has positive cosmological constant Expanding form of Universe
I ₁₅	01050705: <i>Boundary</i>	4d Universe Bounds an anti-de Sitter universe Beings here are mathematically identical to beings in an anti-de Sitter universe Difficult to calculate/analyse

Identifier	Space name	Elements (paraphrased for concision)
I ₁₆	O1O2O6I6: <i>Fish</i>	Inside Goldfish bowl Think Holographic fish are illusion Mathematically equivalent to Holographic fish
I ₁₇	O31O114O: <i>Goldfish bowl</i>	Fish's Reality
I ₁₈	O1O5O1O2: <i>Reality</i>	[Elements are solely schematic or from Knowledge for the Discourse]
I ₁₉	O1O4O7O3: <i>Hologram (of fish)</i>	Image replica of original Fish Mirror movement of original Fish Exist as flat surface Think Fish are illusion Mathematically equivalent to Fish
I ₂₀	O1O4O7O5: <i>Baryon</i>	Has mass and properties
I ₂₂	O1O6O4O2: <i>Information problem</i>	Concerns loss of information when Objects are placed in a black hole
I ₂₃	O1O6O4O2: <i>Mathematical problems</i>	Such as the masses of quarks or the Information problem Easier to calculate in 5d

g: Relationship Diagram

[Overleaf.]



h: Blended Spaces

Identifier	Space name	Source	Elements (paraphrased for concision)
B ₁	01011006: <i>Black Hole 2</i>	I ₉ I ₁₀ I ₁₁ I ₁₂	<p>Has spherical event horizon</p> <p>Contains Information proportional to the surface area of this event horizon</p> <p>Not Book-like</p> <p>Not standard Object</p> <p>Can be judged by surface area</p>
B ₂	X: <i>Duality</i>	I ₁₃ I ₁₅	<p><i>Identity</i> between 4d and 5d universes</p> <p>Beings are <i>reciprocally</i> mathematically identical between 4d and 5d universes</p>
B ₃	X: <i>Analogy</i> [=01050102: <i>Reality 2</i>]	I ₁₆ I ₁₈ I ₁₉ B ₂	<p>3d Fish and 2d Holographic fish are <i>reciprocally</i> mathematically identical</p> <p>3d and 2d identities are joined</p> <p>There is no way to tell which is 'real'</p>

Identifier	Space name	Source	Elements (paraphrased for concision)
B ₄	01060402: <i>Information problem 2</i>	I ₂₂ B ₁ B ₂ B ₃	<p>Information in a black hole is proportional to its surface area</p> <p>Surface areas are dimensional junctures</p> <p>Dimensions joined by a dimensional juncture are mathematically identical</p> <p>When putting information into a black hole, it is passed to a further mathematically identical dimension</p> <p>Information is not lost by passage through a black hole</p>
B ₅	01060402: <i>Mathematical problems 2</i>	I ₂₃ B ₂ B ₄	<p>Such as the masses of quarks or the Information problem</p> <p>Easier in 5d</p> <p>Easier in different dimensions</p>

Identifier	Space name	Source	Elements (paraphrased for concision)
B _X	X: <i>Prediction</i> [=OIOIIIO: <i>Universe</i> 2]	I ₁ I ₃ B ₂ B ₃	<p>Possibly a hologram</p> <p>Any dimensional universe encodes all information necessary to produce an alternate-dimensionality universe</p> <p>Beings are <i>reciprocally</i> mathematically identical between universes</p> <p>There is no way to tell which is 'real'</p> <p>Our universe may be of a different order of dimensionality than we think and there is no way to tell</p>

i: Blending Commentary

B₁ creates a reconceptualization of I₉ OIOIIIO06: *Black hole* following the counterfactual (the key Vital Relation is Disanalogy; I₉ is not I₁₁ or I₁₂). B₂ takes elements of I₁₃ OIIIOIO: *Anti-de Sitter universe* and the I₁₅ OIO50705: *Boundary universe* and fuses them in accordance with CIT practice (identity in one input space is fused into a reciprocal dual identity in the blend); it is named by the cataphoric space referent duality in the sixth paragraph. B₃ analogizes B₂ to the Fish sequence, describing Reality as it is, not as one set of fish see it (the original mental space represented the point of view of the 3d fish). B₄ solves the I₂₂ OIO60402: *Information problem* by taking into account the new conceptualization of black holes (B₁), the B₂ X: *Duality*, and the analogous B₃ of the goldfish, and has the emergent property of surface areas being dimensional

junctures (cf B_2 and B_3) leading to conclusions when ‘running’ the blend which are in boldface above. With B_2 and B_4 in mind, B_5 generalizes I_{23} beyond the specific 5d element in-text. Finally, B_X is the *prediction* cataphorically referred to at the start of the text extract as a reconceptualization of I_1 010110: *Universe*, where the junction of that space, the I_3 01040703: *Hologram* space and the B_2 X: *Duality* and B_3 01050102: *Reality 2 (goldfish)* blends give a generalization of holograms from B_3 to a general statement, leading to emergent structure which is the discourse conclusion.

APPENDIX II.3: *WARPED PASSAGES*

a: Full Text

In *Roald Dahl's Charlie and the Chocolate Factory*, Willy Wonka introduced visitors to his 'Wonkavator.' In his words, 'An elevator can only go up and down, but a Wonkavator goes sideways and slantways and longways and backways and frontways and squareways and any other ways that you can think of..' Really, what he had was a device that moved in any direction, so long as it was a direction in the three dimensions we know. It was a nice, imaginative idea.

However, the Wonkavator didn't really go any way 'you can think of.' Willy Wonka was remiss in that he neglected extra-dimensional passages. Extra dimensions are other directions entirely. They are hard to describe, but they may be easier to understand by analogy.

In 1884, to explain the notion of extra dimensions, the English mathematician Edwin A. Abbott wrote a novel called *Flatland*. It takes place in a fictitious two-dimensional universe – the Flatland of the title – where two-dimensional beings (of various geometric shapes) reside. Abbott shows us why Flatlanders, who live their whole lives in two dimensions – on a table top, for example – are as mystified by three dimensions as people in our world are by the idea of four.

For us, more than three dimensions requires a stretch of the imagination, but in Flatland three dimensions are beyond its inhabitants' comprehension. Everyone thinks it is obvious that the universe holds no more than their two perceived dimensions. Flatlanders are as insistent about this as most people here are about three.

The book's narrator, A. Square (the namesake of the author, Edwin A), is introduced to the reality of a third dimension. In the first stage of his education, while he is still confined to Flatland, he watches a three-dimensional

sphere travel vertically through his two-dimensional world. Because A. Square is confined to Flatland, he sees a series of disks that increase and then decrease in size, which are slices of the sphere as it passes through A. Square's plane (see Figure 6).

This is initially perplexing to the two-dimensional narrator, who has never imagined more than two dimensions and has never contemplated a three-dimensional object like a sphere. It is not until A. Square has been lifted out of Flatland into the surrounding three-dimensional world that he can truly imagine a sphere. From his new perspective, he recognizes the sphere as the shape made by gluing together the two-dimensional slices he witnessed. Even in his two-dimensional world, A. Square could have plotted the disks he sees as a function of time (as in Figure 6) to construct the sphere. But it wasn't until his trip through a third dimension opened his eyes that he fully comprehended the sphere and its third spatial dimension.

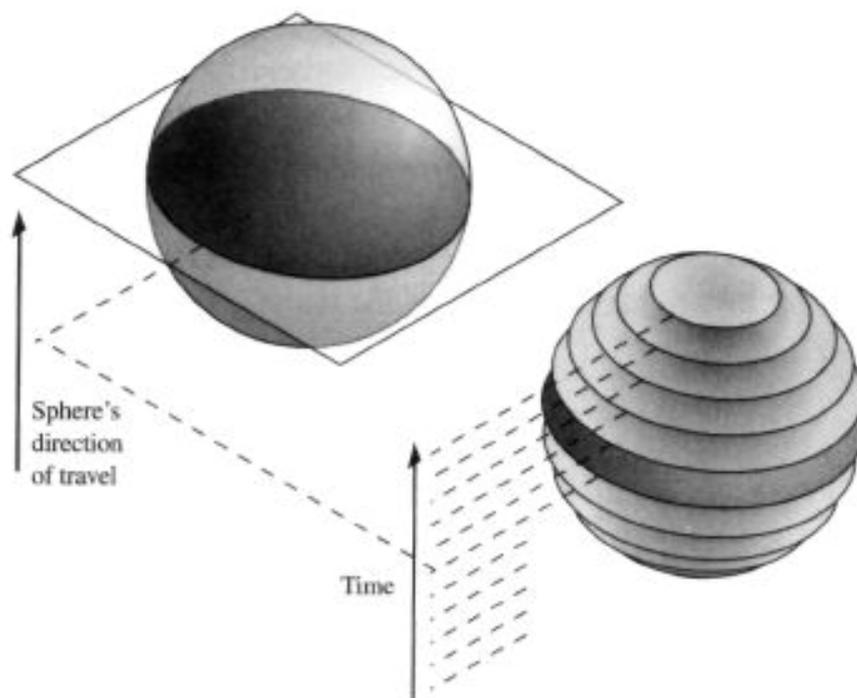


Figure 6. If a sphere passes through a plane, a two-dimensional observer would see a disk. The sequence of disks that the observer sees over time comprises the sphere.

By analogy, we know that if a *hypersphere* (a sphere with four spatial dimensions) were to pass through our universe, it would appear to us as a time sequence of three-dimensional spheres that increase, then decrease, in size. Unfortunately, we don't have the opportunity to journey through an extra dimension. We will never see a static hypersphere in its entirety. Nonetheless, we can make deductions about how objects look in spaces of different dimensions – even dimensions that we don't see. We can confidently deduce that our perception of a hypersphere passing through three dimensions would look like a series of three-dimensional spheres.

As another example, let's imagine the construction of a hyper-cube – a generalization of a cube to more than three dimensions. A line segment of one dimension consists of two points connected by a straight, one-dimensional line. We can generalize this in two dimensions to a square by putting one of these one-dimensional line segments above another and connecting them with two additional segments. We can generalize further in three dimensions to a cube, which we can construct by placing one two-dimensional square above the other and connecting them with four additional squares, one on each edge of the original squares (see Figure 7).

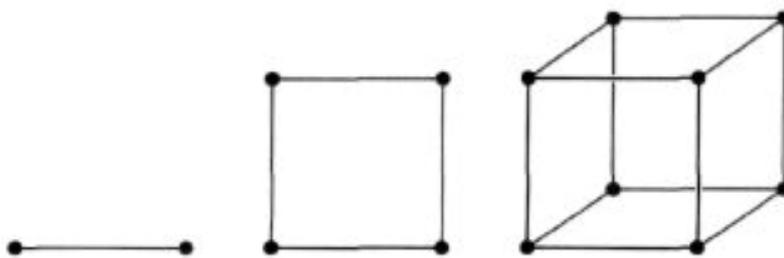


Figure 7. How we put together lower-dimensional objects to make higher-dimensional ones. We connect two points to make a line segment, two line segments to make a square, two squares to make a cube, and (not pictured since it's too difficult to draw) two cubes to make a hypercube.

We can generalize in four dimensions to a hypercube, and in five dimensions to something for which we don't yet have a name. Even though we three-

dimensional mortals have never seen these two objects, we can generalize the procedure that worked in lower dimensions. To construct a hypercube (also known as a tesseract), put one cube above the other, and connect them by adding six additional cubes, connecting the faces of the two original cubes. This construction is an abstraction and difficult to draw, but that doesn't make the hypercube any less real.

In high school, I spent a summer at math camp (which was far more entertaining than you might think), where we were shown a film version of *Flatland*. At the end, the narrator, in a delightful British accent, tried futilely to point to the third dimension that was inaccessible to Flatlanders, saying, 'Upward, not Northward.' Unfortunately, we have the same frustration if we try to point to a fourth spatial dimension, a passage. But just as Flatlanders didn't see or travel through the third dimension, even though it existed in Abbott's story, our not having yet seen another dimension doesn't mean there is none. So although we haven't yet observed or traveled through such a dimension, the subtext throughout *Warped Passages* will be, 'Not Northward, but Forward along a passage.' Who knows what exists that we haven't yet seen?

b: Annotated Text

[Circumstance omitted] 'An elevator can only go up and down, but a **Wonkavator** goes sideways and slantways and longways and backways and frontways and squareways and any other ways that you can think of...' Really, what he had was a *device* that moved in any direction, so long as it was a direction in the three **dimensions** we know. [Metarepresentation omitted.]

However, the *Wonkavator* didn't really go *any way* 'you can think of.' Willy Wonka was remiss in that he neglected extra-dimensional passages >. Extra *dimensions* are other *directions* entirely. [Rhetoric omitted.] they may be easier to understand by *analogy* >.

[Circumstance omitted] to explain the notion of *extra dimensions*, [circumstantial agency omitted] a fictitious two-dimensional universe – the *Flatland* of the title – where two-dimensional beings (of various geometric shapes) reside. [Circumstantial agency omitted] Flatlanders, who live their whole lives in two dimensions – on a **table top**, for example – are as mystified by three dimensions as people in **our world** are by the idea of four.

[Circumstance omitted], but in *Flatland* three *dimensions* are beyond its inhabitants' comprehension. *Everyone* thinks it is obvious that *the universe* holds no more than their two perceived *dimensions*. Flatlanders are as insistent about this as most people here are about three.

The book's narrator, **A. Square** [circumstantial agency omitted], is introduced to the reality of a third *dimension*. In the first stage of his education, while he is still confined to *Flatland*, he watches a three-dimensional sphere travel vertically through his *two-dimensional world*. Because *A. Square* is confined to *Flatland*, he sees a series of **disks** that increase and then decrease in size, which are slices of the sphere as it passes through *A. Square's / plane* (see Figure 6).

This is initially perplexing to the two-dimensional *narrator*, who has never imagined more than two *dimensions* and has never contemplated a three-dimensional object like a *sphere*. It is not until *A. Square* has been lifted out of *Flatland* into the surrounding *three-dimensional world* that he can truly imagine a *sphere*. From his new perspective, he recognizes the *sphere* as the shape made by gluing together the two-dimensional slices he witnessed. Even in his two-dimensional *world*, *A. Square* could have plotted the disks he sees as a function of time (as in Figure 6) to construct the *sphere*. But it wasn't until his trip through a third *dimension* opened his eyes that he fully comprehended the *sphere* and its third spatial *dimension*.

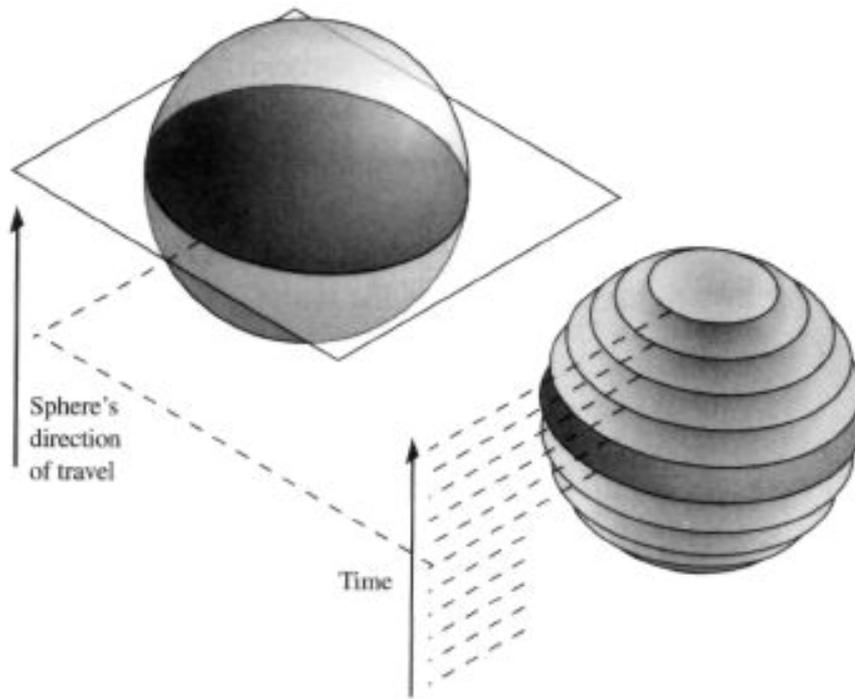


Figure 6. If a sphere passes through a plane, a two-dimensional observer would see a disk. The sequence of disks that the observer sees over time comprises the sphere.

[Rhetoric omitted] we know that if a **hypersphere** (a sphere with four spatial *dimensions*) were to pass through our *universe*, it would appear to us as a time sequence of three-dimensional spheres that increase, then decrease, in size. Unfortunately, we don't have the opportunity to journey through an extra *dimension*. We will never see a static *hypersphere* in its entirety. Nonetheless, we can make deductions about how *objects* look in spaces of different *dimensions* – even *dimensions* that we don't see. We can confidently deduce that our perception of a *hypersphere* passing through three *dimensions* would look like a series of three-dimensional *spheres*.

As another example, let's imagine the construction of a **hyper-cube** [*sic*] – a generalization of a cube to more than three *dimensions*. A **line** segment of one *dimension* consists of two points connected by a straight, one-dimensional *line*. We can generalize this in two dimensions to a **square** by putting one of these

one-dimensional line segments above another and connecting them with two additional segments. We can generalize further in three dimensions to a **cube**, which we can construct by placing one two-dimensional *square* above the other and connecting them with four additional *squares*, one on each edge of the original *squares* (see Figure 7).

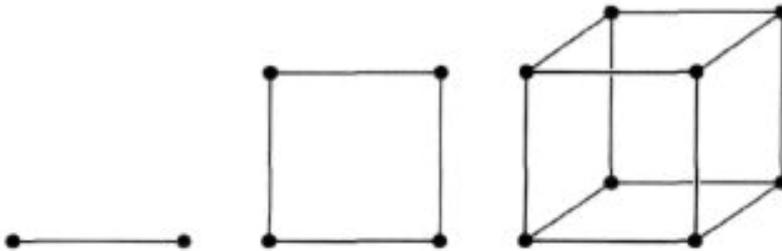


Figure 7. How we put together lower-dimensional objects to make higher-dimensional ones. We connect two points to make a line segment, two line segments to make a square, two squares to make a cube, and (not pictured since it's too difficult to draw) two cubes to make a hypercube.

We can generalize in four dimensions to a *hypercube*, and in five dimensions to **something** for which we don't yet have a name. [Rhetoric omitted.] To construct a *hypercube* (also known as a tesseract), put one *cube* above the other, and connect them by adding six additional *cubes*, connecting the faces of the two original *cubes*. [Rhetoric omitted.]

[Circumstance omitted] the third *dimension* that was inaccessible to *Flatlanders*, saying, 'Upward, not Northward.' Unfortunately, we have the same frustration if we try to point to a fourth spatial *dimension*, a *passage*. But just as *Flatlanders* didn't see or travel through the third *dimension*, even though it existed in Abbott's *story*, our not having yet seen another *dimension* doesn't mean there is none. So although we haven't yet observed or traveled through such a *dimension*, the subtext throughout *Warped Passages* will be, 'Not Northward, but Forward along a passage.' [Rhetoric omitted.]

c: Grouped Mental Spaces

Identifier	Space name	Semantic grouping	Textual grouping
I ₁	03090202: <i>Elevator</i>	030902 TRAVEL	WONKA
I ₂	03090202: <i>Wonkavator</i>	030902 TRAVEL	WONKA
I ₃	01050702: <i>Dimension</i>	0105 EXISTENCE	WONKA/FLATLAND
I ₄	010110: <i>Flatland Universe</i>	010110 UNIVERSE	FLATLAND
I ₅	03101139: <i>Table</i>	031011 EQUIPMENT	FLATLAND
I ₆	010110: <i>Universe</i>	010110 UNIVERSE	FLATLAND
I ₇	X01050703: <i>A. Square</i>	0105 EXISTENCE	FLATLAND
I ₈	01050703: <i>Sphere</i>	0105 EXISTENCE	FLATLAND
I ₉	01050703: <i>Disc</i>	0105 EXISTENCE	FLATLAND
I ₁₀	01050703: <i>Hypersphere</i>	0105 EXISTENCE	HYPER-
I ₁₁	01050703: <i>Hypercube</i>	0105 EXISTENCE	HYPER-
I ₁₂	01050705: <i>Line</i>	0105 EXISTENCE	HYPER-
I ₁₃	01050703: <i>Square</i>	0105 EXISTENCE	HYPER-
I ₁₄	01050703: <i>Cube</i>	0105 EXISTENCE	HYPER-
I ₁₅	X01050703: <i>5d 'Something'</i>	0105 EXISTENCE	HYPER-

d: Mental Spaces and Elements

Identifier	Space name	Elements (paraphrased for concision)
I ₁	03090202: <i>Elevator</i>	Movement in some directions
I ₂	03090202: <i>Wonkavator</i>	Movement in three directions Directions are dimensions
I ₃	01050702: <i>Dimension</i>	Humans know of three Extra dimensions are extra directions

Identifier	Space name	Elements (paraphrased for concision)
I ₄	010110: <i>Flatland Universe</i>	Two dimensions Contain 2d beings Beings are geometric shapes Beings cannot comprehend 3d Beings cannot understand 'upward' direction
I ₅	03101139: <i>Table</i>	Two-dimensional
I ₆	010110: <i>Universe</i>	Three dimensions Beings cannot comprehend 4d Beings cannot understand 4d 'forward along a passage' direction
I ₇	X01050703: <i>A. Square</i>	Flatlander Is educated in narrative
I ₈	01050703: <i>Sphere</i>	Non-Flatlander 3d
I ₉	01050703: <i>Disc</i>	Flatland representation of sphere Change in size Slices of sphere
I ₁₀	01050703: <i>Hypersphere</i>	4d Would appear as time sequence of 3d spheres
I ₁₁	01050703: <i>Hypercube</i>	4d 4d generalization of cube
I ₁₂	01050705: <i>Line</i>	1d Connects two points
I ₁₃	01050703: <i>Square</i>	2d Connects four lines 2d generalization of line

Identifier	Space name	Elements (paraphrased for concision)
I ₁₄	01050703: <i>Cube</i>	3d Connects six squares 3d generalization of square
I ₁₅	X01050703: <i>5d</i> <i>'Something'</i>	5d 5d generalization of hypercube

e: Relationship Diagram

[Overleaf.]

I₁₂ Line
I₁₃ Square
I₁₄ Cube

I₁₅ 5d Something
I₁₁ Hypercube

I₁₀ Hypersphere

B_I

I₆ Universe

I₅ Table

I₈ Sphere

I₄ Flatland Universe

I₉ Disc

I₇ A. Square

I₃ Dimension

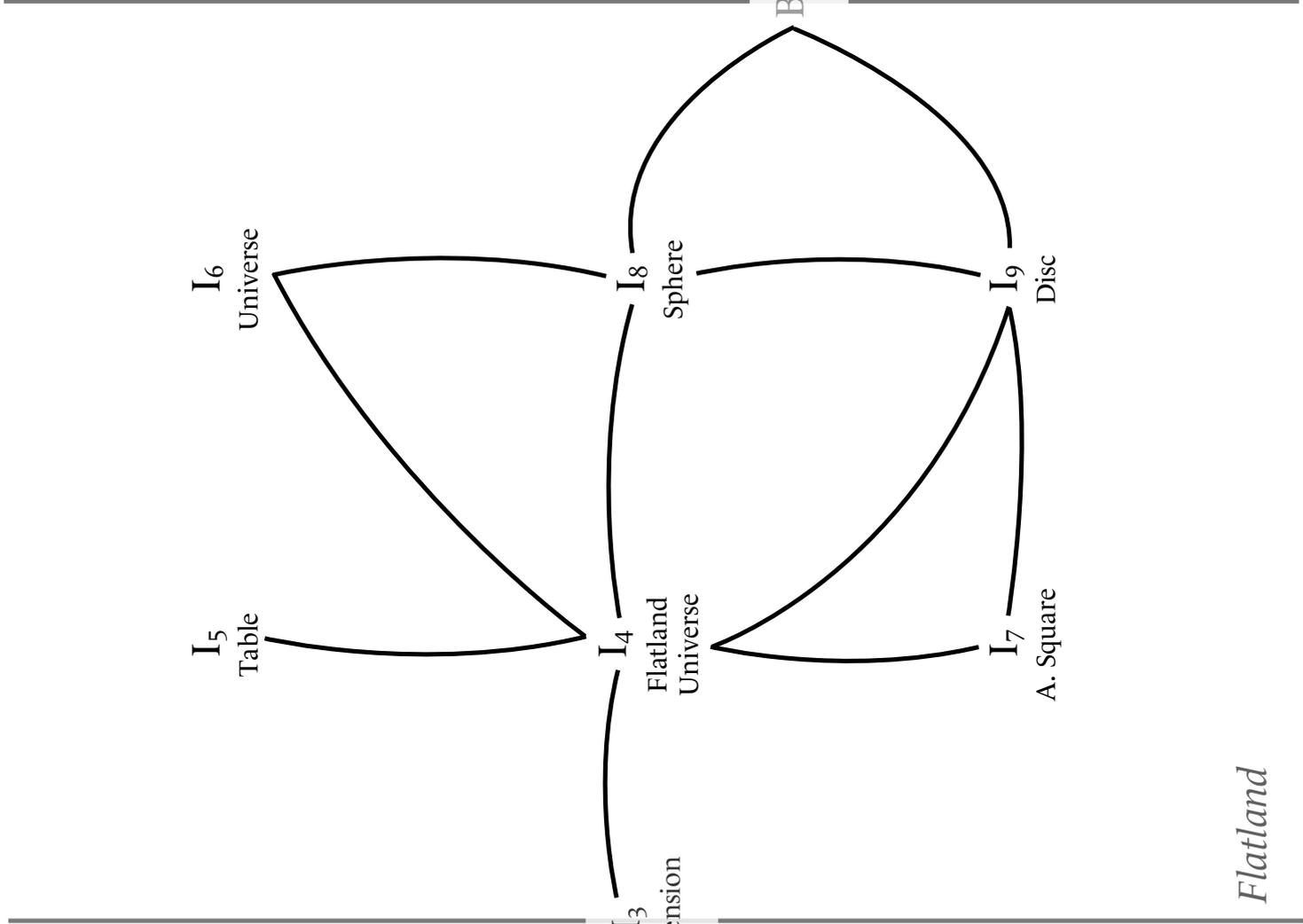
I₁ Elevator

I₂ Wonkavator

Hyper-

Flatland

Wonka



f: Blended Spaces

Identifier	Space name	Source	Elements (paraphrased for concision)
B _I	<i>X: Higher-Dimensional Movement as Change in Intersection</i>	I ₈ I ₉ I ₁₀	<p>Sphere/disc <i>identity</i></p> <p>Representation of 3d object in lower dimension is a change of 2d size</p> <p>Dimensional intersection of a higher dimensional object is represented by perceived change of the intersection</p>
B _X	<i>X: Analogy [Dimensional Passage]</i>	I ₄ I ₆ B _I	<p>2d beings cannot comprehend 3d ‘upward’ direction</p> <p>3d beings cannot comprehend 4d ‘forward along a passage’ direction</p> <p>Dimensional intersection of a higher dimensional object is represented by perceived change of the intersection</p> <p>There exist higher dimensions which are represented by new directions along which there is perceived change</p>

g: Blending Commentary

B_I is the key blend in this passage, which is achieved through the identity of the 2d disc and the 3d sphere (I₈ I₉ respectively) requiring an explanation for the perceived change, which constitutes the emergent property of the blend (a generalization from the textually-signalled 2d/3d). B_I also achieves the

necessary textual coherence between I_8/I_9 and I_{10} . The eventual blend, the putative explanatory purpose of the text, also explicates the book's title; $B_X X$: *Analogy [Dimensional Passage]*, the analogy pre-enacted at the start of the extract, establishes the convention of discussing higher-order dimensionality as a new direction ('Not Northward, but Forward along a passage').

APPENDIX II.4: *THE FABRIC OF THE COSMOS*

a: Full Text

In addition to the difficulties Kaluza-Klein theory encountered in trying to describe the microworld, there was another reason scientists were hesitant about the approach. Many found it both arbitrary and extravagant to postulate a hidden spatial dimension. It is not as though Kaluza was led to the idea of a new spatial dimension by a rigid chain of deductive reasoning. Instead, he pulled the idea out of a hat, and upon analyzing its implications discovered an unexpected link between general relativity and electromagnetism. Thus, although it was a great discovery in its own right, it lacked a sense of inevitability. If you asked Kaluza and Klein why the universe had five spacetime dimensions rather than four, or six, or seven, or 7,000 for that matter, they wouldn't have had an answer much more convincing than 'Why not?'

More than three decades later, the situation changed radically. String theory is the first approach to merge general relativity and quantum mechanics; moreover, it has the potential to unify our understanding of all forces and all matter. But the quantum mechanical equations of string theory don't work in four spacetime dimensions, nor in five, six, seven, or 7,000. Instead, for reasons discussed in the next section, the equations of string theory work only in ten spacetime dimensions – nine of space, plus time. String theory *demand*s more dimensions.

This is a fundamentally different kind of result, one never before encountered in the history of physics. Prior to strings, no theory said anything at all about the number of spatial dimensions in the universe. Every theory from Newton to Maxwell to Einstein assumed that the universe had three space dimensions, much as we all assume the sun will rise tomorrow. Kaluza and Klein proffered a challenge by suggesting that there were four space dimensions, but this

amounted to yet another assumption – a different assumption, but an assumption nonetheless. Now, for the first time, string theory provided equations that predicted the number of space dimensions. A calculation – not an assumption, not a hypothesis, not an inspired guess – determines the number of space dimensions according to string theory, and the surprising thing is that the calculated number is not three, but nine. String theory leads us, *inevitably*, to a universe with six extra space dimensions and hence provides a compelling, ready-made context for invoking the ideas of Kaluza and Klein.

The original proposal of Kaluza and Klein assumed only one hidden dimension, but it's easily generalized to two, three, or even the six extra dimensions required by string theory. For example, in Figure 12.8a we replace the additional circular dimension of Figure 12.7, a one dimensional shape, with the surface of a sphere, a two-dimensional shape (recall from the discussion in Chapter 8 that the surface of a sphere is two dimensional because you need two pieces of information – like latitude and longitude on the earth's surface – to specify a location). As with the circle, you should envision the sphere tacked on to every point of the usual dimensions, even though in Figure 12.8a, to keep the image clear, we draw only those that lie on the intersections of grid lines. In a universe of this sort, you would need a total of five pieces of information to locate a position in space: three pieces to locate your position in the big dimensions (street, cross street, floor number) and two pieces to locate your position on the sphere (latitude, longitude) tacked on at that point. Certainly, if the sphere's radius were tiny – billions of times smaller than an atom – the last two pieces of information wouldn't matter much for comparatively large beings like ourselves. Nevertheless, the extra dimension would be an integral part of the ultramicroscopic makeup of the spatial fabric. An ultramicroscopic worm would need all five pieces of information and, if we include time, it would need six pieces of information in order to show up at the right dinner party at the right time.

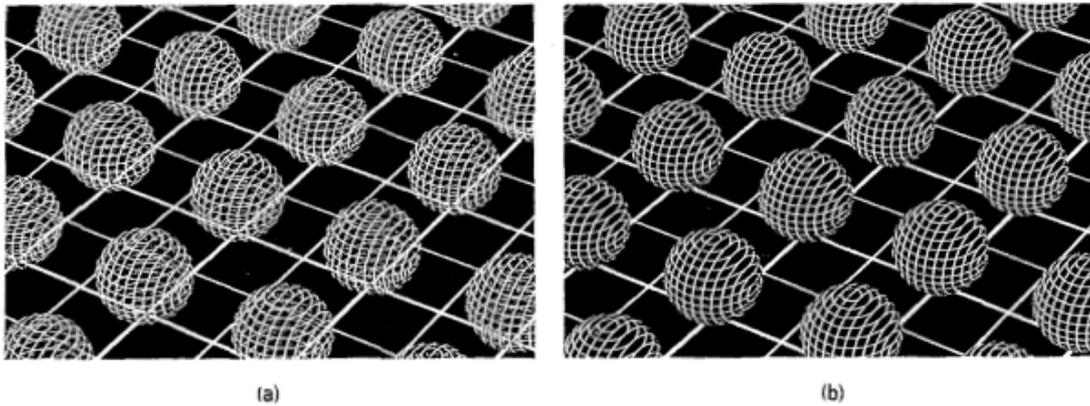


Figure 12.8 A close-up of a universe with the three usual dimensions, represented by the grid, and (a) two curled-up dimensions, in the form of hollow spheres, and (b) three curled-up dimensions in the form of solid balls.

Let's go one dimension further. In Figure 12.8a, we considered only the surface of the spheres. Imagine now that, as in Figure 12.8b, the fabric of space also includes the interior of the spheres – our little Planck-sized worm can burrow into the sphere, as ordinary worms do with apples, and freely move throughout its interior. To specify the worm's location would now require six pieces of information: three to locate its position in the usual extended spatial dimensions, and three more to locate its position in the ball tacked on to that point (latitude, longitude, depth of penetration). Together with time, this is therefore an example of a universe with *seven* spacetime dimensions.

Now comes a leap. Although it is impossible to draw, imagine that at every point in the three extended dimensions of everyday life, the universe has not one extra dimension as in Figure 12.7, not two extra dimensions as in Figure 12.8a, not three extra dimensions as in Figure 12.8b, but six extra space dimensions. I certainly can't visualize this and I've never met anyone who can. But its meaning is clear. To specify the spatial location of a Planck-sized worm in such a universe requires *nine* pieces of information: three to locate its

position in the usual extended dimensions and six more to locate its position in the curled-up dimensions tacked on to that point. When time is also taken into account, this is a ten-spacetime-dimensional universe, as required by the equations of string theory. If the extra six dimensions are curled up small enough, they would easily have escaped detection.

b: Annotated Text

[Circumstantial agency omitted.] **String theory** is the first approach to merge **general relativity** and **quantum mechanics**; moreover, it has the potential to unify our understanding of all forces and all matter. But the quantum mechanical equations of *string theory* don't work in four spacetime **dimensions**, nor in five, six, seven, or 7,000. Instead, for reasons discussed in the next section, the *equations* of *string theory* work only in ten spacetime *dimensions* – nine of space, plus time. *String theory* demands more *dimensions*.

[Circumstance omitted.] Prior to *strings*, no theory said anything at all about the number of spatial *dimensions* in the *universe*. [Circumstance omitted] that there were four space *dimensions*, but this amounted to yet another assumption – a different assumption, but an assumption nonetheless. Now, for the first time, *string theory* provided equations that predicted the number of space *dimensions*. A *calculation* – not an assumption, not a hypothesis, not an inspired guess – determines the number of space *dimensions* according to *string theory*, and the surprising thing is that the calculated number is not three, but nine. *String theory* leads us, inevitably, to a **universe** with six extra space *dimensions* [circumstance omitted].

[Circumstantial agency omitted.] in Figure 12.8a we replace the additional circular *dimension* of Figure 12.7, a one dimensional **shape**, with the surface of a **sphere**, a two-dimensional shape (recall from the discussion in Chapter 8 that the surface of a *sphere* is two dimensional because you need two pieces of information – like **latitude** and **longitude** on the earth's surface – to specify a

location). As with the *circle*, you should envision the *sphere* tacked on to every point of the usual *dimensions* [rhetoric omitted]. In a *universe* of this sort, you would need a total of five pieces of **information** to locate a *position* in space: three *pieces* to locate your *position* in the big *dimensions* (street, cross street, floor number) and two pieces to locate your *position* on the *sphere* (*latitude, longitude*) tacked on at that point. Certainly, if the sphere's radius were tiny – billions of times smaller than an atom – [rhetoric omitted]. Nevertheless, the extra *dimension* would be an integral part of the ultramicroscopic makeup of the spatial *fabric*. An ultramicroscopic **worm** would need all five pieces of *information* and, if we include time, it would need six pieces of *information* in order to show up at the right dinner **party** at the right time.

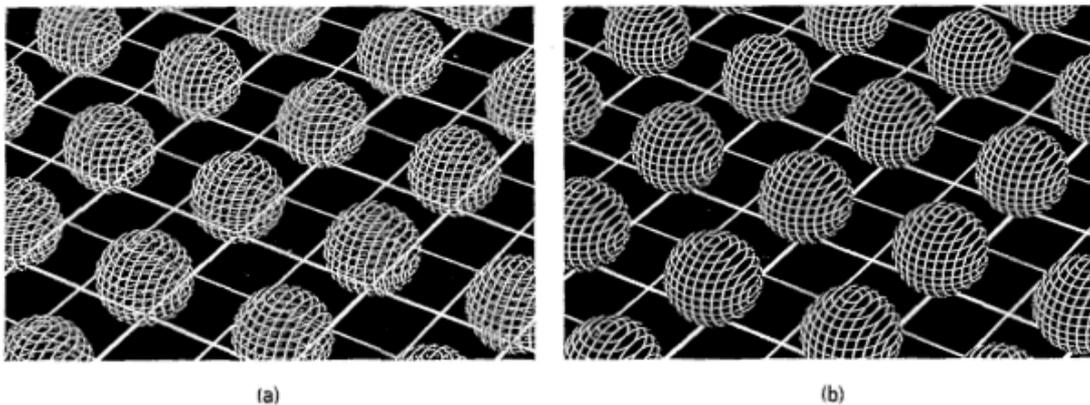


Figure 12.8 A close-up of a universe with the three usual dimensions, represented by the grid, and (a) two curled-up dimensions, in the form of hollow spheres, and (b) three curled-up dimensions in the form of solid balls.

Let's go one *dimension* further. In Figure 12.8a, we considered only the surface of the *spheres*. Imagine now that, as in Figure 12.8b, the *fabric of space* also includes the interior of the *spheres* – our little Planck-sized *worm* can burrow into the *sphere*, as ordinary *worms* do with **apples**, and freely move throughout its *interior*. To specify the worm's *location* would now require six pieces of

information: three to locate its position in the usual extended spatial *dimensions*, and three more to locate its position in the *ball* tacked on to that point (*latitude, longitude, depth of penetration*). Together with time, this is therefore an example of a *universe* with seven spacetime *dimensions*.

[Rhetoric omitted] imagine that at every point in the three extended *dimensions* of everyday life, the *universe* has not one extra *dimension* as in Figure 12.7, not two extra *dimensions* as in Figure 12.8a, not three extra *dimensions* as in Figure 12.8b, but six extra space *dimensions*. [Circumstance omitted.] To specify the spatial *location* of a Planck-sized *worm* in such a universe requires nine pieces of *information*: three to locate its position in the usual extended *dimensions* and six more to locate its *position* in the curled-up *dimensions* tacked on to that *point*. When *time* is also taken into account, this is a ten-spacetime-dimensional *universe*, as required by the *equations* of *string theory*. If the extra six *dimensions* are curled up small enough, they would easily have escaped detection.

c: Grouped Mental Spaces

Identifier	Space name	Semantic grouping	Textual grouping
I ₁	01040705: <i>String theory</i>	010407 PHYSICS	THEORY
I ₂	01040708: <i>Relativity</i>	010407 PHYSICS	THEORY
I ₃	01040707: <i>Quantum theory</i>	010407 PHYSICS	THEORY
I ₄	01050702: <i>Dimension</i>	0105 EXISTENCE	THEORY
I ₅	010110: <i>Universe</i>	0101 UNIVERSE	DIMENSIONS
I ₆	01050703: <i>Shape</i>	0105 EXISTENCE	DIMENSIONS
I ₇	01050703: <i>Sphere</i>	0105 EXISTENCE	DIMENSIONS
I ₈	010102: <i>Geodetic reference</i>	0101 UNIVERSE	DIMENSIONS
I ₉	01050704: <i>Location</i>	0105 EXISTENCE	DIMENSIONS
I ₁₀	01020613: <i>Worm</i>	01020613 WORM	DIMENSIONS

Identifier	Space name	Semantic grouping	Textual grouping
I ₁₁	03110205: <i>Dinner party</i>	03110205 PARTY	DIMENSIONS
I ₁₂	01020413: <i>Apple</i>	01020413 APPLE	DIMENSIONS

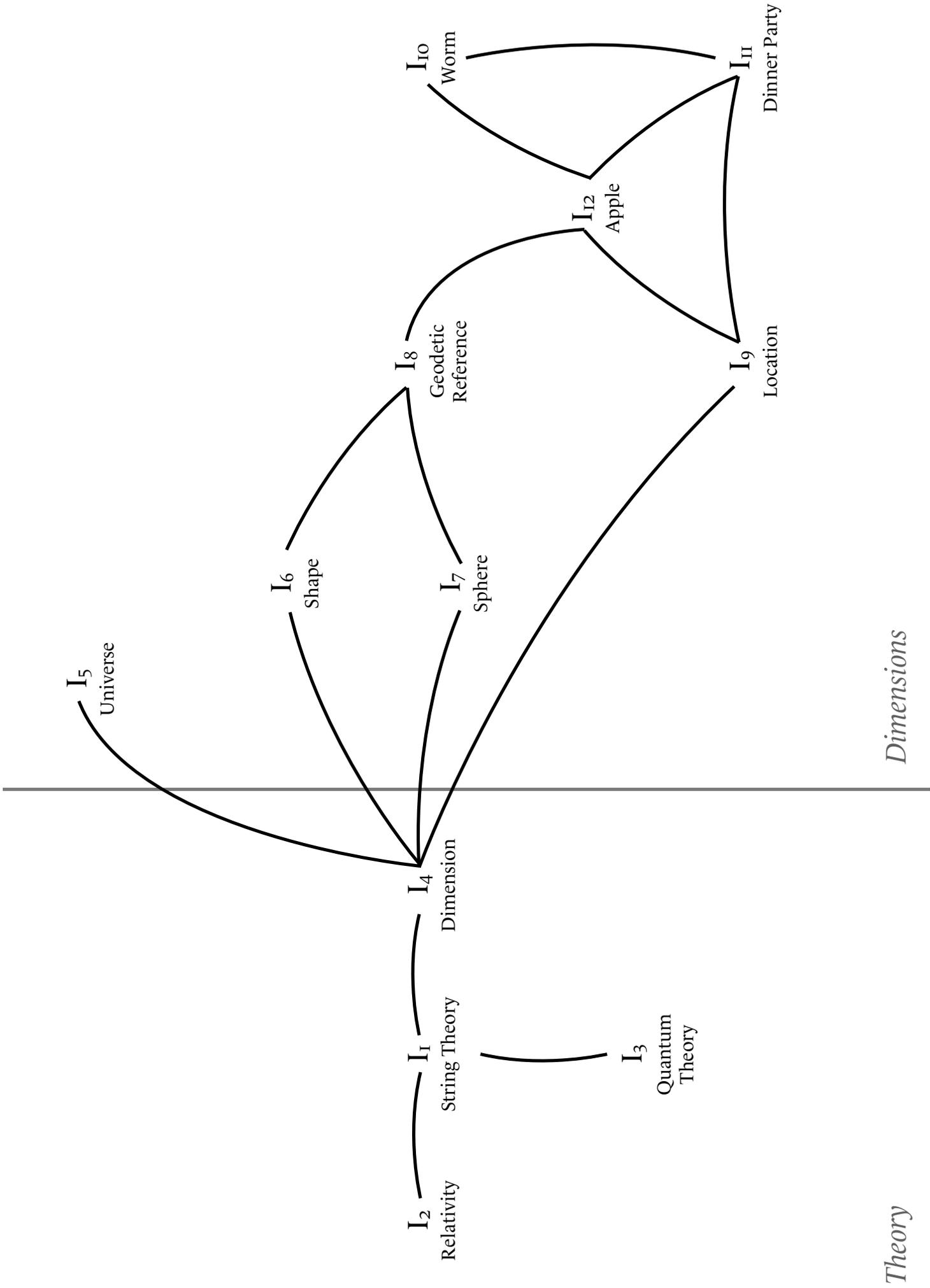
d: Mental Spaces and Elements

Identifier	Space name	Elements (paraphrased for concision)
I ₁	01040705: <i>String theory</i>	Has equations Those equations only work in 10 dimensions Nine of these are space, one time Previous theories had fewer dimensions Dimensions calculated using equations
I ₂	01040708: <i>Relativity</i>	[Elements are solely schematic or from Knowledge for the Discourse]
I ₃	01040707: <i>Quantum theory</i>	Has equations of string theory
I ₄	01050702: <i>Dimension</i>	Can be spatial or temporal Can escape detection by being curled
I ₅	010110: <i>Universe</i>	Has 10 dimensions
I ₆	01050703: <i>Shape</i>	2d
I ₇	01050703: <i>Sphere</i>	3d Tacked on to every point of the usual dimensions
I ₈	010102: <i>Geodetic reference</i>	2 of which reference Location on 2d Shape's surface, eg latitude and longitude 3 of which reference Location in 3d space, eg street, cross street, floor number
I ₉	01050704: <i>Location</i>	Elements constitute dimensionality
I ₁₀	01020613: <i>Worm</i>	Ultramicroscopic Travelling to Dinner party

Identifier	Space name	Elements (paraphrased for concision)
I ₁₁	03110205: <i>Dinner party</i>	Exists within dimensionality Has Location
I ₁₂	01020413: <i>Apple</i>	Solid version of Sphere Need an additional Geodetic reference, depth of penetration, to have Location

e: Relationship Diagram

[Overleaf.]



f: Blended Spaces

Identifier	Space name	Source	Elements (paraphrased for concision)
B ₁	01050702: <i>Dimension 2</i>	I ₄ I ₉ I ₁₂	Can be spatial or temporal Locational elements constitute dimensionality Additional locational elements require higher dimensionality
B ₂	010102: <i>Geodetic reference 2</i>	B ₂ I ₈	Additional locational elements require higher dimensionality n locational references require n dimensions
B _X	010110: <i>Universe 2</i>	I ₁ I ₅ B ₂	String theory has nine dimensions of space and one of time The universe has ten dimensions n dimensions require n locational references The universe exists in 10 locational references

g: Blending Commentary

B₁ is created as a reanalysis of I₄ from the input in I₉ and I₁₂, where location and dimensionality are linked. The I₁₂ 010204I3: *Apple* space gives the link between extra information and dimensionality, as does I₉, implicitly. B₁ therefore makes this link explicit, as does B₂ for the link between this dimensionality and geodetic references. Finally, B_X takes I₁ 01040705: *String theory* and I₅ 010110: *Universe* alongside B₂ to redefine the universe as one with locational references,

making sense of the final sentence 'If the extra six dimensions are curled up small enough, they would easily have escaped detection'.

APPENDIX II.5: *THE MUSIC OF THE PRIMES*

Background. *The Music of the Primes* uses some more technical vocabulary than the other texts in this analysis, here briefly explained. *Prime numbers*, a subset of normal numbers that can only be exactly divided by themselves or the number one⁵⁴, are unpredictable by any mathematical means. Unlike many other number series, primes cannot be found by looking at previous primes, but instead must be found empirically (testing each number to discover if it has exact divisors other than itself and one). The ability to predict the distribution of primes is a highly sought-after goal in mathematics, and in 1859 a hypothesis was put forth by Bernhard Riemann that primes can be predicted by looking at where the graph of a particular function (known as the *Riemann zeta function*) crosses the x-axis, or zero line, of the graph. The zero function uses *imaginary numbers*, numbers which do not exist in real terms but are instead formed by the impossible result of $\sqrt{-1}$. This result, i , has the property of i^2 being equal to -1 , but as no number can be squared to form -1 (1×1 is 1 , and -1×-1 is also 1), the result is termed *imaginary*. Accordingly, the extract below discusses both imaginary numbers and the graph of the Riemann zeta function.

a: Full Text

Riemann had found a passageway from the familiar world of numbers into a mathematics which would have seemed utterly alien to the Greeks who had studied prime numbers two thousand years before. He had innocently mixed imaginary numbers with his zeta function and discovered, like some mathematical alchemist, the mathematical treasure emerging from this admixture of elements that generations had been searching for. He had crammed his ideas into a ten-page paper, but was fully aware that his ideas would open up radically new vistas on the primes.

Riemann's ability to unleash the full power of the zeta function stems from critical discoveries he made during his Berlin years and in his later doctoral

⁵⁴ 'Exactly divided' here means to divide without leaving a remainder. 6 can be exactly divided by 3 (the result is a whole number, 2) but cannot be exactly divided by 4 (the result is 1.5, not a whole number).

studies in Göttingen. What had so impressed Gauss while he was examining Riemann's thesis was the strong geometric intuition that the young mathematician showed when he was feeding functions with imaginary numbers. After all, Gauss had capitalised on his own private mental picture to map out these imaginary numbers before he dismantled the conceptual scaffolding. The starting point for Riemann's theory of these imaginary functions had been Cauchy's work, and for Cauchy a function was defined by an equation. Riemann had now added the idea that even if the equation was the starting point, it was the geometry of the graph defined by the equation that really mattered.

The problem is that the complete graph of a function fed with imaginary numbers is not something that is possible to draw. To illustrate his graph, Riemann needed to work in four dimensions. What do mathematicians mean by a fourth dimension? Those who have read cosmologists such as Stephen Hawking might well reply 'time'. The truth is that we use dimensions to keep track of anything we might be interested in. In physics there are three dimensions for space and a fourth dimension for time. Economists who wish to investigate the relationship between interest rates, inflation, unemployment and the national debt can interpret the economy as a landscape in four dimensions. As they trek uphill in the direction interest rates [*sic*], they will be exploring what happens to the economy in the other directions. Although we can't actually draw a picture of this four-dimensional model of the economy, it is still a landscape whose hills and troughs we can analyse.

For Riemann, the zeta function was similarly described by a landscape that existed in four dimensions. There were two dimensions to keep track of the coordinates of the imaginary numbers being fed into the zeta function. The third and fourth dimensions could then be used to record the two coordinates describing the imaginary number output by the function.

The trouble is that we humans exist in three spatial dimensions and so cannot rely on our visual world for a perception of this new ‘imaginary graph’.

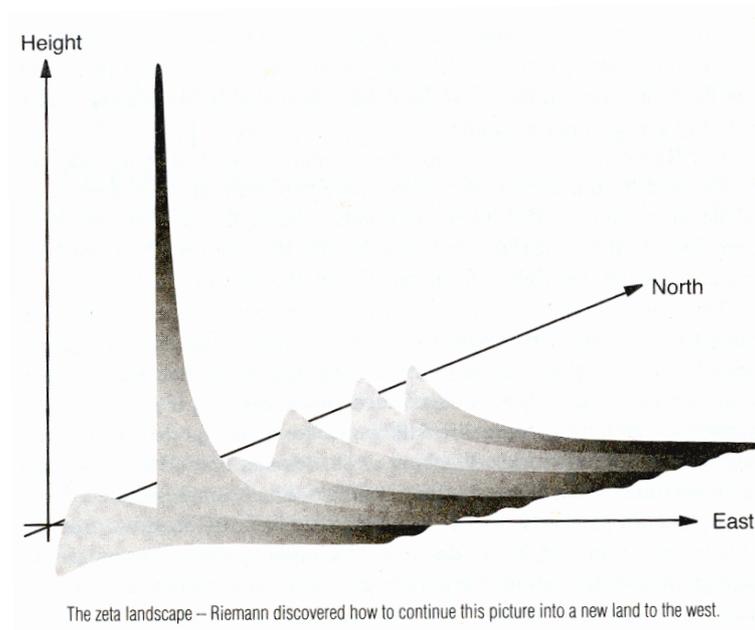
Mathematicians have used the language of mathematics to train their mind’s eye to help them ‘see’ such structures. But if you lack such mathematical lenses, there are still ways to help you to grasp these higher-dimensional worlds.

Looking at shadows is one of the best ways to understand them. Our shadow is a two-dimensional picture of our three-dimensional body. From some perspectives the shadow provides little information, but from side-on, for example, a silhouette can give us enough information about the person in three dimensions for us to recognise their face. In a similar way, we can construct a three-dimensional shadow of the four-dimensional landscape that Riemann built using the zeta function which retains enough information for us to understand Riemann’s ideas.

Gauss’s two-dimensional map of imaginary numbers charts the numbers that we shall feed into the zeta function. The north-south axis keeps track of how many steps we take in the imaginary direction, whilst the east west axis charts the real numbers. We can lay this map out flat on a table. What we want to do is to create a physical landscape situated in the space above this map. The shadow of the zeta function will then turn into a physical object whose peaks and valleys we can explore.

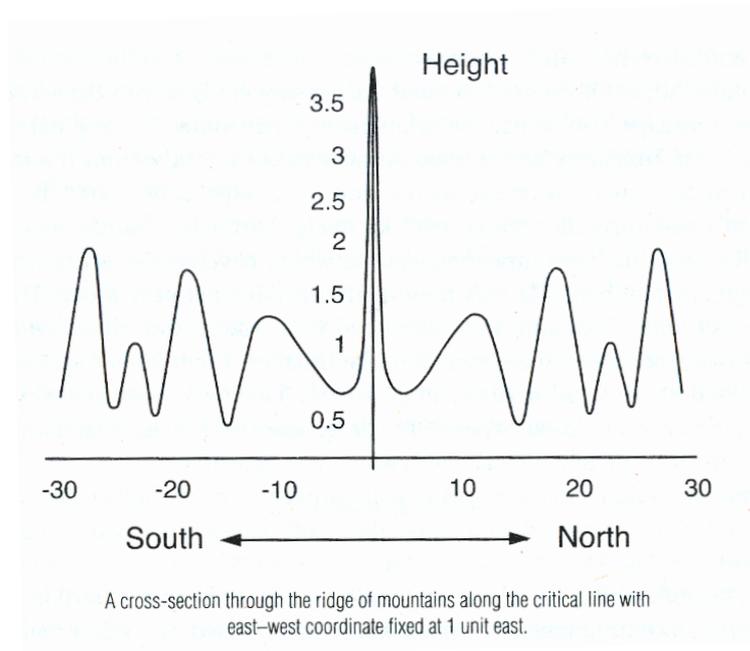
The height above each imaginary number on the map should record the result of feeding that number into the zeta function. Some information is inevitably lost in the plotting of such a landscape, just as a shadow shows very limited detail of a three-dimensional object. By turning this object we get different shadows which reveal different aspects of the object. Similarly, we have a number of choices for what to record as the height of the landscape above each imaginary number in the map on the table top. There is, however, one choice of shadow which retains enough information to allow us to understand Riemann’s revelation. It is a perspective that helped Riemann in his journey

through this looking-glass world. So what does this particular three-dimensional shadow of the zeta function look like?



As Riemann began to explore this landscape, he came across several key features. Standing in the landscape and looking towards the east, the zeta landscape levelled out to a smooth plane 1 unit high above sea level. If Riemann turned round and started walking west, he saw a ridge of undulating hills running from north to south. The peaks of these hills were all located above the line that crossed the east-west axis through the number 1. Above this intersection at the number 1 there was a towering peak which climbed into the heavens. It was, in fact, infinitely high. As Euler had learned, feeding the number 1 into the zeta function gives an output which spirals off to infinity. Heading north or south from this infinite peak, Riemann encountered other peaks. None of these peaks, however, were infinitely high. The first peak occurred at just under 10 steps north at the imaginary number $1 + (9.986\dots)i$ and was only about 1.4 units high.

If Riemann turned the landscape around and charted the cross-section of the hills running along this north-south divide through the number ι , it would look something like this:



b: Annotated Text

[Circumstantial agency omitted] a **mathematics** which [circumstantial agency omitted] studied **prime numbers** [circumstantial agency omitted] mixed **imaginary numbers** with his **zeta function** and discovered, like some mathematical **alchemist**, the mathematical **treasure** emerging from this **admixture** of **elements** that generations had been searching for.

[Circumstantial agency omitted] his ideas would open up radically new **vistas** on the *primes*.

[Circumstances omitted] a **function** was defined by an **equation**.

[Circumstantial agency omitted] even if the *equation* was the starting point, it was the geometry of the **graph** defined by the *equation* that really mattered.

The problem is that the complete *graph* of a *function* fed with *imaginary numbers* is not something that is possible to draw. To illustrate his *graph*, Riemann needed to work in four **dimensions**. [Rhetoric and circumstances omitted] *dimensions* to keep track of anything we might be interested in. In **physics** there are three *dimensions* for space and a fourth dimension for time. [Circumstantial agency omitted] the relationship between interest rates, inflation, unemployment and the national debt can interpret the **economy** as a **landscape** in four *dimensions*. As they trek uphill in the direction interest rates [*sic*], they will be exploring what happens to the *economy* in the other directions. [Circumstances omitted] it is still a *landscape* whose hills and troughs we can analyse.

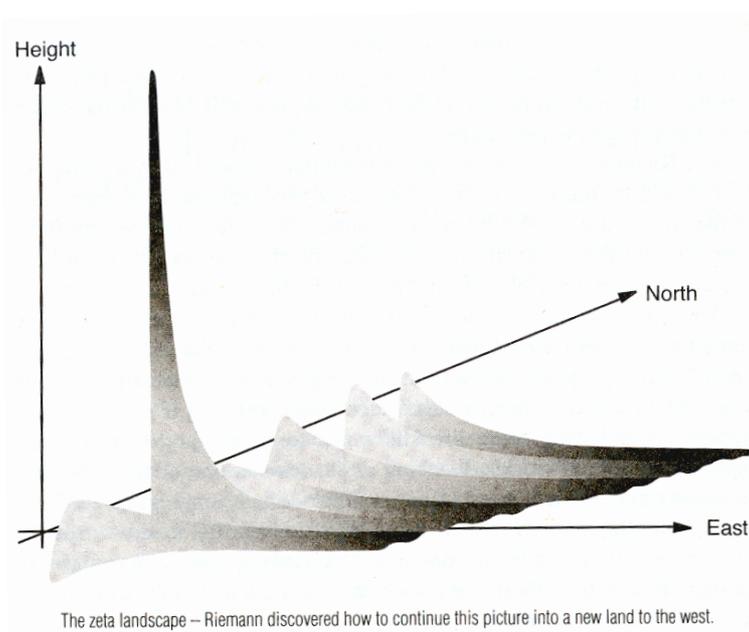
[Circumstantial agency omitted] the *zeta function* was similarly described by a *landscape* that existed in four *dimensions*. There were two *dimensions* to keep track of the coordinates of the *imaginary numbers* being fed into the *zeta function*. The third and fourth *dimensions* could then be used to record the two coordinates describing the *imaginary number* output by the *function*.

[Rhetoric omitted] to grasp these higher-dimensional **worlds**. Looking at **shadows** is one of the best ways to understand *them*. Our *shadow* is a two-dimensional **picture** of our three-dimensional **body**. From some perspectives the *shadow* provides little *information*, but from side-on, for example, a *silhouette* can give us enough information about the *person* in three dimensions for us to recognise their face. In a similar way, we can construct a **three-dimensional shadow** of the four-dimensional *landscape* [circumstantial agency omitted] built using the *zeta function* which retains enough information [circumstance omitted].

[Circumstantial agency omitted] two-dimensional **map** of *imaginary numbers* charts the numbers that we shall feed into the *zeta function*. The north-south axis keeps track of how many steps we take in the imaginary direction, whilst the east west axis charts the **real numbers**. We can lay this *map* out flat on a **table**.

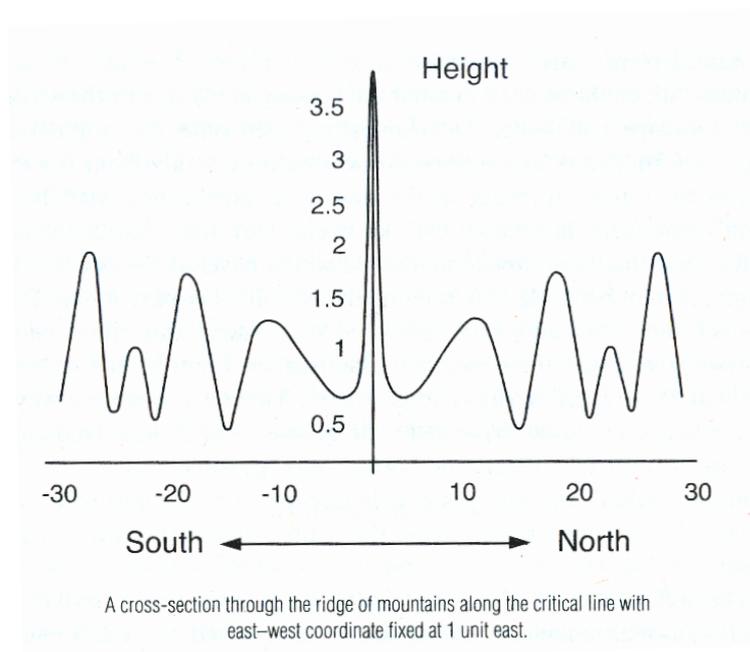
What we want to do is to create a physical *landscape* situated in the space above this *map*. The *shadow* of the *zeta function* will then turn into a physical object whose peaks and valleys we can explore.

The **height** above each *imaginary number* on the *map* should record the result of feeding that number into the *zeta function*. Some information is inevitably lost in the plotting of such a *landscape*, just as a *shadow* shows very limited detail of a three-dimensional *object*. By turning this *object* we get different *shadows* which reveal different aspects of the *object*. Similarly, we have a number of choices for what to record as the *height* of the *landscape* above each *imaginary number* in the *map* on the *table top*. There is, however, one choice of *shadow* which retains enough information to allow us to understand [circumstantial agency omitted] *revelation* >. It is a perspective that helped Riemann in his journey through this looking-glass *world*. So what does this particular three-dimensional *shadow* of the *zeta function* look like?



As Riemann began to explore this *landscape*, he came across several key features. Standing in the *landscape* and looking towards the *east*, the *zeta landscape* levelled out to a smooth plane 1 unit high above sea level. If Riemann turned round and started walking *west*, he saw a ridge of undulating hills running from north to south. The peaks of these hills were all located above the line that crossed the east-west axis through the number 1. Above this intersection at the number 1 there was a towering peak which climbed into the heavens. It was, in fact, infinitely high. As Euler had learned, feeding the number 1 into the *zeta function* gives an output which spirals off to infinity. Heading north or south from this infinite peak, Riemann encountered other peaks. None of these peaks, however, were infinitely high. The first peak occurred at just under 10 steps north at the imaginary number $1 + (9.986\dots)i$ and was only about 1.4 units high.

If Riemann turned the *landscape* around and charted the cross-section of the hills running along this north-south divide through the number 1, it would look something like this:



c: Grouped Mental Spaces

Identifier	Space name	Semantic grouping	Textual grouping
I ₁	01060402: Mathematics	010604 MATHEMATICS	MATHEMATICS
I ₂	01060403: <i>Prime number</i>	010604 MATHEMATICS	MATHEMATICS
I ₃	01060403: <i>Imaginary number</i>	010604 MATHEMATICS	MATHEMATICS
I ₄	01060407: <i>Zeta function</i>	010604 MATHEMATICS	MATHEMATICS
I ₅	010401: <i>Alchemy</i>	0104 MATTER	ALCHEMY
I ₆	03101315: <i>Treasure</i>	031013 COMMERCE	ALCHEMY
I ₇	01060705: <i>Admixture</i>	0106 RELATIVE PROPERTIES	ALCHEMY
I ₈	01040217: <i>Elements</i>	0105 EXISTENCE	ALCHEMY
I ₉	01030710: <i>Vista</i>	0101 THE WORLD	LANDSCAPE
I ₁₀	01060407: <i>Function</i>	010604 MATHEMATICS	LANDSCAPE
I ₁₁	01060407: <i>Equation</i>	010604 MATHEMATICS	LANDSCAPE
I ₁₂	01060412: <i>Graph</i>	010604 MATHEMATICS	LANDSCAPE
I ₁₃	01050702: <i>Dimension</i>	0105 EXISTENCE	LANDSCAPE
I ₁₄	010407: <i>Physics</i>	0104 MATTER	LANDSCAPE
I ₁₅	03101320: <i>Economy</i>	031013 COMMERCE	LANDSCAPE
I ₁₆	01010404: <i>Landscape</i>	0101 THE WORLD	LANDSCAPE
I ₁₇	010110: <i>Higher-dimensional world</i>	0101 THE WORLD	SHADOW
I ₁₈	01040810: <i>Shadow</i>	0104 MATTER	SHADOW
I ₁₉	03110302: <i>Picture</i>	031103 IMAGE	SHADOW

Identifier	Space name	Semantic grouping	Textual grouping
I ₂₀	01050104: <i>Object</i>	0105 EXISTENCE	SHADOW
I ₂₁	01040810: <i>Three-dimensional shadow</i>	0104 MATTER	SHADOW
I ₂₂	01010901: <i>Map</i>	0101 THE WORLD	LANDSCAPE
I ₂₃	01060403: <i>Real numbers</i>	010604 MATHEMATICS	LANDSCAPE
I ₂₄	03101139: <i>Table</i>	031011 EQUIPMENT	LANDSCAPE
I ₂₅	01050702: <i>Height</i>	0105 EXISTENCE	LANDSCAPE
I ₂₆	01010505: <i>Sea level</i>	0101 THE WORLD	LANDSCAPE
I ₂₇	X01050705: <i>I-line</i>	X0105 I-LINE	LANDSCAPE

d: Mental Spaces and Elements

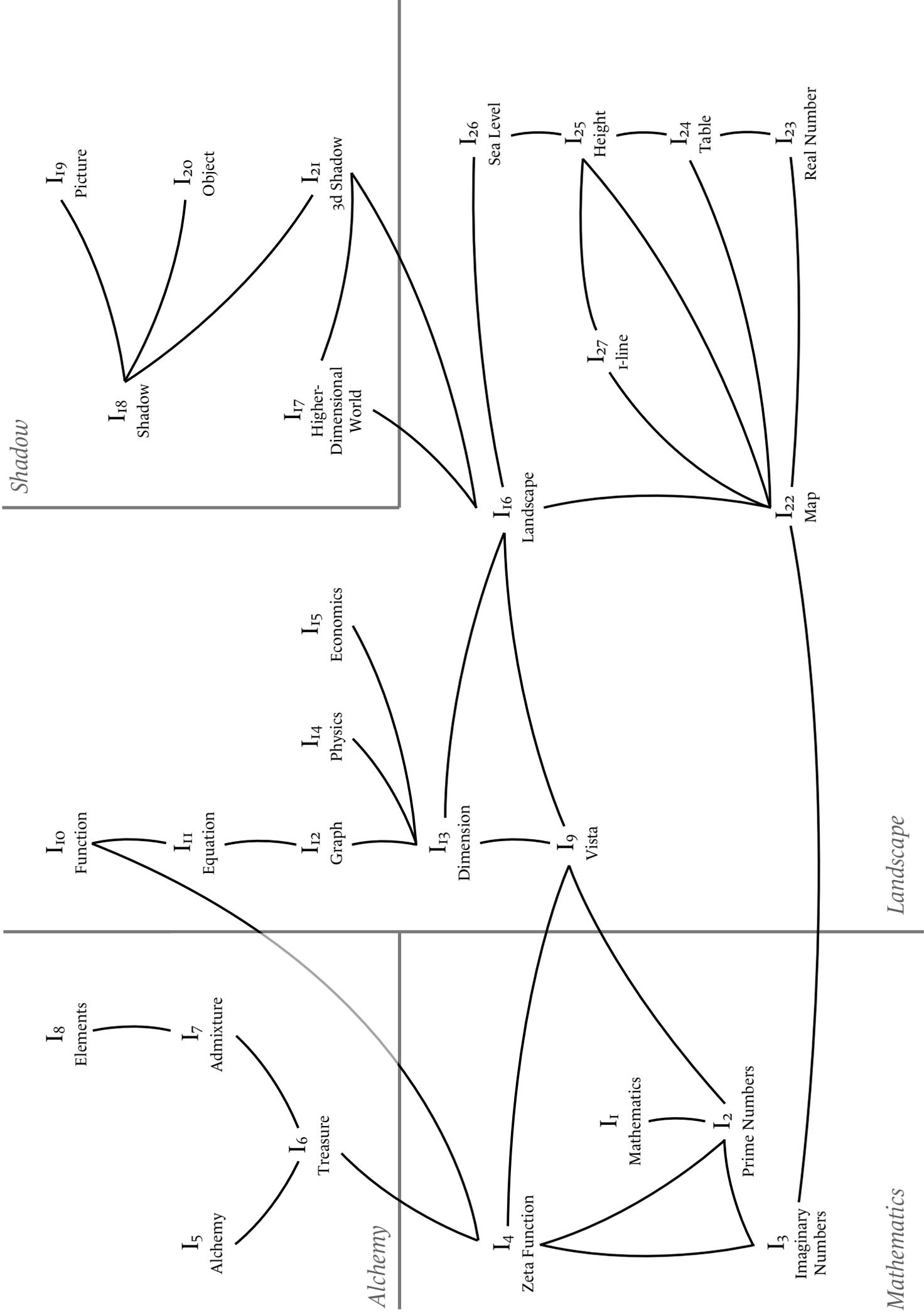
Identifier	Space name	Elements (paraphrased for concision)
I ₁	01060402: <i>Mathematics</i>	[Elements are solely schematic or from Knowledge for the Discourse]
I ₂	01060403: <i>Prime number</i>	[Elements which are schematic or from Knowledge for the Discourse: <i>A natural number, greater than 1, that has no divisors other than 1 and itself</i>]
I ₃	01060403: <i>Imaginary number</i>	[Elements which are schematic or from Knowledge for the Discourse: <i>A number whose square is a Real number less than 0</i>]
I ₄	01060407: <i>Zeta function</i>	Has significance for prime numbers Can be interpreted as a 4d Landscape Two dimensions: coordinates of Imaginary numbers input Two dimensions: coordinates of Imaginary numbers output

Identifier	Space name	Elements (paraphrased for concision)
I ₅	010401: <i>Alchemy</i>	[Elements are solely schematic or from Knowledge for the Discourse]
I ₆	03101315: <i>Treasure</i>	[Elements are solely schematic or from Knowledge for the Discourse]
I ₇	01060705: <i>Admixture</i>	[Elements are solely schematic or from Knowledge for the Discourse]
I ₈	01040217: <i>Elements</i>	[Elements are solely schematic or from Knowledge for the Discourse]
I ₉	01030710: <i>Vista</i>	[Elements are solely schematic or from Knowledge for the Discourse]
I ₁₀	01060407: <i>Function</i>	Defined by an Equation
I ₁₁	01060407: <i>Equation</i>	Defines a Function
I ₁₂	01060412: <i>Graph</i>	Has geometry Defined by an Equation
I ₁₃	01050702: <i>Dimension</i>	Tracks elements of interest Humans live in a three-dimensional world
I ₁₄	010407: <i>Physics</i>	Four-dimensional
I ₁₅	03101320: <i>Economy</i>	Can be interpreted as a 4d Landscape Has multiple elements (such as interest rates, inflation, unemployment and the national debt)
I ₁₆	01010404: <i>Landscape</i>	Four-dimensional Has hills, troughs, directions
I ₁₇	010110: <i>Higher-dimensional world</i>	4d
I ₁₈	01040810: <i>Shadow</i>	Collapses 3d into 2d Contains variable information depending on perspective Perspectives can give information on a 3d Object Loses information
I ₁₉	03110302: <i>Picture</i>	A 2d representation of a 3d Object
I ₂₀	01050104: <i>Object</i>	3d

Identifier	Space name	Elements (paraphrased for concision)
I ₂₁	01040810: <i>Three-dimensional shadow</i>	Collapses 4d into 3d
I ₂₂	01010901: <i>Map</i>	2d Charts Imaginary numbers Has two axes; north-south and east-west Input to Zeta function
I ₂₃	01060403: <i>Real numbers</i>	[Elements which are schematic or from Knowledge for the Discourse: <i>Any number which can be found on the number line</i>]
I ₂₄	03101139: <i>Table</i>	[Elements are solely schematic or from Knowledge for the Discourse]
I ₂₅	01050702: <i>Height</i>	[Elements are solely schematic or from Knowledge for the Discourse]
I ₂₆	01010505: <i>Sea level</i>	[Elements are solely schematic or from Knowledge for the Discourse]
I ₂₇	X01050705: <i>1-line</i>	Crosses the east-west axis of a Landscape Map at number 1

e: Relationship Diagram

[Overleaf.]



f: Blended Spaces

Identifier	Space name	Source	Elements (paraphrased for concision)
B ₁	01060407: <i>Zeta function 2</i>	I ₄ I ₆	Has significance for prime numbers Treasure Valuable function/valuable significance
B ₂	01040810: <i>Three-dimensional shadow 2</i>	I ₁₆ I ₁₇ I ₁₈	Has hills, troughs, directions Collapses 4d into 3d Contains variable information depending on perspective Perspectives can give information on a 4d entity Loses information from original 4d
B ₃	01010404: <i>Landscape 2</i>	I ₉ I ₁₃ I ₁₆ I ₂₂ I ₂₆ B ₂	Has hills, troughs, directions Tracks elements of interest Collapses 4d into 3d Charts Imaginary numbers Has north-south and east-west axes Has hills, troughs, directions Perspectives can give information on a 4d entity
B ₄	01010901: <i>Map 2</i>	I ₃ I ₂₂ I ₂₃ I ₂₅	2d Has two axes; north-south (Imaginary numbers) and east-west (Real numbers) Has space for height above

Identifier	Space name	Source	Elements (paraphrased for concision)
B _X	X: <i>Reimann's landscape</i>	I ₂₆ I ₂₇ B ₃ B ₄	<p>Has hills, troughs, directions, Sea level, and a 1-line</p> <p>1-line crosses the east-west axis in height of this landscape at number 1</p> <p>Tracks 3d elements of interest</p> <p>Represents 4d</p> <p>Has three axes; north-south (Imaginary numbers), east-west (Real numbers), height as result of Zeta function</p> <p>Loses information from original 4d</p>

g: Blending Commentary

B₁ straightforwardly takes the alchemy/treasure comparison and makes explicit that the zeta function is a valuable result. B₂ to B₄ are redefinitions of previous spaces: B₂ concludes the shadow textual group by integrating major information from that group into a single blend, which profiles the nature of a Three-dimensional shadow of a 4d entity; B₃ takes B₂ and the neighbouring input spaces of the landscape analogy to redefine the 01010404: *Landscape* space into one which takes account of the higher-order dimensionality it describes; and B₄ integrates the measures of the map's axes with the concept of height above the map. The full arch-blend, B_X, has been titled X: *Reimann's landscape*, on account of its dominant structural input, and explicitly combines the 01010404: *Landscape 2* and 01010901: *Map 2* blends with the final pieces of the analogy (01010505: *Sea level* and X01050705: *1-line*). It has emergent structure

which gives the nature of the core analogy of this extract; that this landscape of the Zeta function has three axes and is a lossy representation of the 4d function itself.