LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

TRINITY COLLEGE
0 5 AUG 2005
LIBRARY DUBLIN
THOSIS 7676

Methods for Meaningful Text Representation and Comparison

Ann Devitt

Thesis submitted for the Degree of Doctor of Philosophy

Department of Computer Science

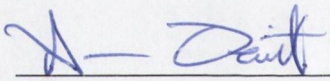
University of Dublin

Trinity College

14 June 2005

Declaration

I hereby declare that this thesis is entirely my own work and that it has not been submitted as an exercise for a degree at any other university and that I agree that the Library may lend or copy the thesis upon request.



June 14, 2005

Ann Devitt

Abstract

This thesis investigates methods of text representation and comparison using WordNet as a real world knowledge base. The research is carried out within the framework of lexical cohesive theory which defines the cohesive structure of text built through the connections and interactions of its constituent lexical items. The model of cohesive structure explored in this thesis is a network of interrelated lexical items.

This thesis presents Storylines, a system to construct a network model from free text. The Storylines model is applied to the task of text differentiation: similarity estimation and differential content extraction for pairs of texts. Storylines uses the WordNet knowledge base as its source of real world knowledge. However, the use of WordNet brings with it the attendent problem of the variant data sparsity of hand-coded knowledge bases. In order to address this issue, empirical investigations were carried out to define a metric of concept specificity and a weighting algorithm for relation types in WordNet. These parameters, along with text-based measures of cohesive devices are exploited in the text differentiation algorithm.

In order to validate system performance on the text differentiation tasks, separate gold standards of human similarity and difference judgments are derived from experiments. A thorough performance evaluation of the system is carried out using these datasets.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Carl Vogel, for first introducing me to this field in 1996 but mainly for his consistently insightful ideas and unfailing encouragement over the course of this research.

I also acknowledge the Senior Lecturer's Broad Curriculum Initiative for the Teaching Fellowship which allowed me to return to Ireland to undertake this research.

I would also like to thank the other members of the Computational Linguistic Group, in particular, Tim Fernando for his support and good chats and most especially Jennifer Foster who it has been a pleasure to teach with over the last three years and who, with her generosity of spirit, has been a great help and support during the last few months.

I am also very grateful to Michael Lee at the University of Adelaide in Australia for providing me with data for testing my system and for very informative and prompt responses to any questions. Bree Treacy at RTÉ interactive was also very kind in allowing me any resources I required to build the corpus of RTÉ news texts.

Outside the college domain, I have been blessed with wonderful friends who have given me great support and kept me grounded with great chats and good laughs. It would have been no fun without you! Thank you.

To my darling family, Mum, Dad, Marie, Jim, Seamus, Daniela and beau-

tiful Sofia, who have all helped me in hundreds of ways, I count myself very lucky to be among you.

Finally, to the wonderful Gareth Bennett without whom I would not have completed this thesis, thank you for making the last three years the loveliest years of my life!

Contents

1 Introduction **1**

1.1 Thesis Aims 2

1.2 Thesis Outline 4

2 Meaningful Text Representation: Theory and Applications **7**

2.1 Introduction 8

2.2 Text *vs.* Untext 8

2.3 Unstructured representations 10

2.3.1 Term Frequency Vector Models 12

2.3.2 Explicitly Stochastic models 13

2.3.3 Latent Semantic Analysis 14

2.4 Coherence: Structured Representations 16

2.4.1 Rhetorical Structure Theory 17

2.4.2 Grosz and Sidner (G&S) 19

2.4.3 Bridging the Gap between RST and G&S 20

2.4.4 Applications of Theories of Coherence 22

2.5 Theory of Cohesion: Structured Representations 23

2.5.1 Reference 25

2.5.2 Substitution and Ellipsis 26

2.5.3 Conjunction 27

2.6 Lexical cohesion 28
2.6.1 Lexical Cohesive Devices 30
Reiteration 30
Collocation 30
2.6.2 Structure: Nets and Chains 32
2.6.3 Lexical Chaining Algorithms 33
Morris and Hirst 34
Hirst and St-Onge 35
Barzilay and Elhadad 36
Stairmand 37
Harabagiu and Moldovan 37
2.6.4 Implementations of Cohesion 38
Discourse Analysis 39
Word Sense Disambiguation and Semantic Similarity . 41
Summarization 43
Text Classification and Semantic Orientation 44
2.7 Discussion 45

3 Materials: Knowledge Base and Corpora **48**
3.1 Introduction 49
3.2 WordNet 49
3.2.1 Synonymy and the Synset 51
Synset contents 53
3.2.2 Relational semantics 55
Synonymy and Antonymy 56
Hyponymy / Hypernymy: @ and ~ 56
Meronymy / holonymy: # and % 57
Cross-syntactic category relations 58

3.2.3 Syntactic categories in WordNet 62
Nouns in WordNet 62
Verbs in WordNet 64
Modifiers in WordNet 65
3.2.4 Inadequacies of WordNet 67
3.3 Corpora 68
3.3.1 The RTÉ Corpus 69
RTÉ and RTÉ Interactive 69
The RTÉ Interactive editing process 70
The Texts 70
3.3.2 Australian Corpus 71
3.3.3 Google News Corpus 71
3.4 Summary 72

4 The Topology of WordNet **73**

4.1 Introduction 74
4.2 Topological features 76
4.2.1 Some basic measures 76
4.2.2 Dimensional distribution 77
4.2.3 Multiple inheritance quantified 78
4.2.4 Hierarchy Overlap 81
4.2.5 Branching Factor 82
4.2.6 Depth and Height 84
4.2.7 Coordinate terms: node sisters 87
4.2.8 Clustering coefficients 88
4.2.9 Polysemy in WordNet 90
4.3 Specificity experiment 92
4.3.1 Experiment Goals 92

4.3.2 Methodology 93
Participants 93
Materials 93
Procedure 94
4.3.3 Results 95
4.4 Some Conclusions on WordNet and Node Specificity 97

5 Meaningful Text Representation: an Implementation 99
5.1 Introduction 100
5.2 Aims 101
5.3 From Text to Graph: Basic Algorithm 102
5.4 Module 1: Part of Speech Tagging 104
5.5 Module 2: Term extraction 106
5.6 Module 3: WordNet relation extraction 110
5.6.1 Algorithm overview 111
Synset extraction 111
Relation Extraction 113
5.6.2 Required WordNet relations 115
Relations for Recursion 119
Textual Relations 120
Recursive Relation Extraction Process 121
5.7 Module 4: Graph Construction 124
5.8 Text Representation Data Structures 127
5.8.1 The Node Class: WordNet Synsets 131
5.8.2 NodeInGraph: text-based information 133
5.8.3 Arc: WordNet and Text relations 135
5.8.4 Text: the original text 137
5.8.5 WNDag: the text representation class 137

5.9 Latent information content 140
5.10 Discussion 142

6 Text Comparison: Identifying Similarities and Differences 144
6.1 Introduction 145
6.2 Aims 147
6.3 Complexity of the problem 148
6.4 Basic algorithm 149
6.5 Deriving the intersection and difference graphs 152
6.6 Measuring graph size 155
6.6.1 Features for weight calculation 157
Traversal count 157
Presence in text 158
Polysemy count 161
Relation type 163
Node specificity value 164
6.6.2 System techniques for weight calculations 168
6.7 Determining significance 173
6.7.1 Possibilities for set intersection 174
6.7.2 Baseline similarity 176
6.7.3 Similarity Measures 178
The Dice and Jaccard Coefficients 179
The Cosine coefficient 180
The Asymmetric coefficient 180
6.7.4 Implemented Measures 181
6.8 Identifying differences 182
6.9 Discussion 184

7 Gold Standard Human Evaluation **186**

7.1 Introduction 187

7.2 Motivations 188

7.3 Method 189

7.3.1 Experiment design and the talk-aloud protocol 189

A training page: to have or to have not? 191

The Rating Scale 191

Similarity *versus* relatedness 193

7.3.2 Implementation 194

Internet experiments 194

Recorded Results 195

7.3.3 Participants 195

7.3.4 Materials 196

7.3.5 Procedure 198

Instructions Page 198

Training phase 199

The Experiment 201

7.4 Results 202

7.4.1 Some basic results 202

Participation 202

Time 202

Revisions 203

Some feedback 204

7.4.2 Distribution of similarity ratings 205

7.4.3 Distribution of relatedness ratings 206

7.4.4 Range of ratings used by individuals 210

7.5 Reliability: Inter-respondent agreement 210

7.5.1 The Data 214
7.5.2 Deviation from the central tendency 216
Deviation of users from the mean rating 216
Deviation of users from the mode 217
Deviation of story pairs from the central tendency . . . 219
Some Comments 220
7.5.3 Percent Agreement 220
7.5.4 Kappa Reliability Coefficient 221
Interpreting the coefficient 223
Calculating P(E) 224
Siegel and Castellan's κ coefficient 225
Krippendorff's α 227
7.5.5 Re-analysing the data 232
Correlation of respondent agreement with mean ratings
A regression analysis 232
Correlation of respondent agreement with mean ratings
Curve estimation analysis 236
Implications of the Regression Analysis 236
7.5.6 Cochran Q statistic 238
7.5.7 Euclidean distance in 6-dimensional space 239
7.5.8 Reliability of the Training Data 244
7.5.9 Conclusions 247

8 Results and Evaluation **249**

8.1 Introduction 250
8.2 Three "Gold" Standard Datasets 251
8.2.1 RTÉ Data 251
8.2.2 RTÉ Rel Data 251

8.2.3 Australian Data 252
Australian Similarity Experiment 252
8.2.4 Australian and RTÉ Experiment Differences 252
8.3 Threshold Estimation 256
8.3.1 Results of Threshold Estimation Process 257
Extreme and balanced threshold values 257
Human Data Thresholds 258
Similarity Measure Thresholds 259
8.3.2 Final Evaluation Test Cases 260
8.4 Performance Metrics 263
8.4.1 Graphic representation of results 268
8.5 Similarity Detection Results 269
8.5.1 RTÉ Gold Standard Evaluation Results 270
Test Cases 270
Similarity Measures 271
Techniques 271
8.5.2 Australian Gold Standard Evaluation Results 272
Test Cases 273
Techniques 274
8.5.3 RTÉ Rel Gold Standard Evaluation Results 274
8.5.4 Synthesis of Results 286
Thresholds and Similarity Measures 286
Techniques 287
8.5.5 Discussion of text similarity 291
8.6 Difference Identification Results 293
8.6.1 Google News Difference Identification Experiment . . . 293
Materials 293

Subjects 294
Procedure 294
Basic Results 294
8.6.2 Google News Evaluation Results 294
8.6.3 Google News Experiment Discussion 297
8.7 Discussion 300

9 Future Work and Conclusions **302**
9.1 Introduction 303
9.2 Future Work 304
9.2.1 Lexical primes and seeding the representation 304
9.2.2 Small World Graphs 305
9.2.3 Multiword Expressions and Proper Nouns 306
9.2.4 Adding Logical and Rhetorical Relations 306
9.2.5 Skewed Similarity Experiment 307
9.2.6 Differential Content Output 307
9.2.7 Concept Specificity 308
9.2.8 Similarity *versus* Relatedness 308
9.3 Thesis Contributions 309
9.4 Concluding Remarks 312

A Corpora **334**
A.1 RTE Story topics 335
A.2 RTE Gold Standard Experiment Texts 338
A.3 Australian Gold Standard Experiment Texts 346
A.4 Google News Experiment Texts 356

B WordNet Specificity Experiment Materials **367**

C StoryLines Materials and Source Code **372**

C.1 LT POS tags 373

C.2 Step 2 Term Extraction source code 375

C.3 Step 3 Relation Extraction source code 386

C.4 WordNet Topographical File Derivation Source Code 393

C.5 Input and Output File Sample 400

C.5.1 Step 1 input 400

C.5.2 Step 1 output = Step 2 input 401

C.5.3 Step 2 output = Step 3 input 402

C.5.4 Step 3 output = Step 4 input 406

C.5.5 Step 4 input: WordNet topographical features file . . . 407

D Gold Standard Experiment Materials **409**

D.1 Talk aloud protocol materials 410

D.2 Gold Standard Experiment Screen Shots 434

D.3 Gold Standard Experiment Source Code 437

E Gold Standard Experiment Results **443**

E.1 Group Rating Tables 444

E.2 Group Rating Graphs 460

E.3 Siegel and Castellan Kappa Statistic Table 492

E.4 Euclidean Distance from Matlab Function *pdist* 497

E.4.1 Tests with artificial data: all different 498

Euclidean distance: all 1s 499

Euclidean distance: all 10s 500

E.4.2 Tests with user data 501

E.5 Story-by-Story Matrices 533

E.5.1 Evaluation text filenames 533

E.5.2 Story Pair Matrix 534
E.5.3 Mean Ratings in Story-by-Story Matrix 536
E.5.4 Mode Ratings in Story-by-Story Matrix 538

List of Figures

4.1 Bar chart of synset distribution in top noun hierarchies 77
4.2 Histogram of depth for noun nodes with multiple parentage . . 79
4.3 Nouns: Histogram of Maximum Depth 85
4.4 Nouns: Histogram of Minimum Depth 86
4.5 Nouns: Maximum Height (distance to any leaf node) 88
4.6 Nouns: Minimum Height (distance to any leaf node) 89

5.1 Storylines System Process Flow Chart 103
5.2 Partial hypernym graph for Text1 130
5.3 Path from hypernym graph for Text1 130

6.1 Venn Diagram with Large Set Intersection 174
6.2 Venn Diagram with Small Set Intersection 175
6.3 Venn diagram with no set difference 175
6.4 Venn Diagram with no set intersection 175
6.5 Venn Diagram with Moderate Set Intersection 176
6.6 Venn Diagram with Set Subsumption 176

7.1 Histogram of response times, 20 second intervals 203
7.2 Bar chart of similarity ratings on full 11-point scale 206
7.3 Histogram of similarity ratings on 6-point scale 207
7.4 Bar chart of relatedness ratings on full 11 point scale 208

7.5 Histogram of difference between similarity and relatedness ratings . . . 209
7.6 Similarity Ratings for Group 1 . . . 215
7.7 Mean Similarity Ratings for Group 1 . . . 216
7.8 Scatterplot of mean similarity rating with mean user deviation 218
7.9 Histogram of average respondent difference from mode . . . 219
7.10 Scatterplot of percentage agreement with mean similarity rating 233
7.11 Scatterplot of percentage agreement with mean similarity rating without outlier and with fitted line . . . 233
7.12 Plot of Observed Values, Linear and Quadratic Curve Estimation Models . . . 237
7.13 Group 20 Euclidean Distance Matrix . . . 243
7.14 Group 23 Euclidean Distance Matrix . . . 244

8.1 Australian: Line Plot of Optimum Thresholds for Jaccard, Dice and Asymmetric measures . . . 260
8.2 Australian Extreme: Line Plot of Optimum Thresholds for Jaccard, Dice and Asymmetric measures . . . 261
8.3 RTÉ: Line Plot of Optimum Thresholds for Jaccard, Dice and Asymmetric measures . . . 262
8.4 RTÉ Data. Unfiltered. . . . 277
8.5 RTÉ Data. Binary result when tolerance is 0. . . . 278
8.6 RTÉ Data. Filtered when tolerance is 0. . . . 279
8.7 Australian Data. Unfiltered. . . . 280
8.8 Australian Data. Binary result when tolerance is 0. . . . 281
8.9 Australian Data. Filtered when tolerance is 0. . . . 282
8.10 RTÉ Rel Data. Unfiltered. . . . 283
8.11 RTÉ Rel Data. Binary result when tolerance is 0. . . . 284

8.12 RTÉ Rel Data. Filtered when tolerance is 0. 285

D.1 Experiment Instructions Page Screenshot 434
D.2 Experiment Training Page Screenshot 435
D.3 Sample Experiment Page Screenshot 436

E.1 Similarity Ratings for Group 1 460
E.2 Mean Similarity Ratings for Group 1 460
E.3 Similarity Ratings for Group 2 461
E.4 Mean Similarity Ratings for Group 2 461
E.5 Similarity Ratings for Group 3 462
E.6 Mean Similarity Ratings for Group 3 462
E.7 Similarity Ratings for Group 4 463
E.8 Mean Similarity Ratings for Group 4 463
E.9 Similarity Ratings for Group 5 464
E.10 Mean Similarity Ratings for Group 5 464
E.11 Similarity Ratings for Group 6 465
E.12 Mean Similarity Ratings for Group 6 465
E.13 Similarity Ratings for Group 7 466
E.14 Mean Similarity Ratings for Group 7 466
E.15 Similarity Ratings for Group 8 467
E.16 Mean Similarity Ratings for Group 8 467
E.17 Similarity Ratings for Group 9 468
E.18 Mean Similarity Ratings for Group 9 468
E.19 Similarity Ratings for Group 10 469
E.20 Mean Similarity Ratings for Group 10 469
E.21 Similarity Ratings for Group 11 470
E.22 Mean Similarity Ratings for Group 11 470

E.23 Similarity Ratings for Group 12 471
E.24 Mean Similarity Ratings for Group 12 471
E.25 Similarity Ratings for Group 13 472
E.26 Mean Similarity Ratings for Group 13 472
E.27 Similarity Ratings for Group 14 473
E.28 Mean Similarity Ratings for Group 14 473
E.29 Similarity Ratings for Group 15 474
E.30 Mean Similarity Ratings for Group 15 474
E.31 Similarity Ratings for Group 16 475
E.32 Mean Similarity Ratings for Group 16 475
E.33 Similarity Ratings for Group 17 476
E.34 Mean Similarity Ratings for Group 17 476
E.35 Similarity Ratings for Group 17 477
E.36 Mean Similarity Ratings for Group 18 477
E.37 Similarity Ratings for Group 19 478
E.38 Mean Similarity Ratings for Group 19 478
E.39 Similarity Ratings for Group 20 479
E.40 Mean Similarity Ratings for Group 20 479
E.41 Similarity Ratings for Group 21 480
E.42 Mean Similarity Ratings for Group 21 480
E.43 Similarity Ratings for Group 22 481
E.44 Mean Similarity Ratings for Group 22 481
E.45 Similarity Ratings for Group 23 482
E.46 Mean Similarity Ratings for Group 23 482
E.47 Similarity Ratings for Group 24 483
E.48 Mean Similarity Ratings for Group 24 483
E.49 Similarity Ratings for Group 25 484

E.50 Mean Similarity Ratings for Group 25 484
E.51 Similarity Ratings for Group 26 485
E.52 Mean Similarity Ratings for Group 26 485
E.53 Similarity Ratings for Group 27 486
E.54 Mean Similarity Ratings for Group 27 486
E.55 Similarity Ratings for Group 28 487
E.56 Mean Similarity Ratings for Group 28 487
E.57 Similarity Ratings for Group 29 488
E.58 Mean Similarity Ratings for Group 29 488
E.59 Similarity Ratings for Group 30 489
E.60 Mean Similarity Ratings for Group 30 489
E.61 Similarity Ratings for Group 31 490
E.62 Mean Similarity Ratings for Group 31 490
E.63 Similarity Ratings for Group 32 491
E.64 Mean Similarity Ratings for Group 32 491
E.65 Euclidean Distance Matrix: all different 498
E.66 Euclidean Distance Matrix: all 1s 499
E.67 Euclidean Distance Matrix: all 10s 500
E.68 Group 1 Euclidean Distance Matrix 501
E.69 Group 2 Euclidean Distance Matrix 502
E.70 Group 3 Euclidean Distance Matrix 503
E.71 Group 4 Euclidean Distance Matrix 504
E.72 Group 5 Euclidean Distance Matrix 505
E.73 Group 6 Euclidean Distance Matrix 506
E.74 Group 7 Euclidean Distance Matrix 507
E.75 Group 8 Euclidean Distance Matrix 508
E.76 Group 9 Euclidean Distance Matrix 509

E.77 Group 10 Euclidean Distance Matrix 510
E.78 Group 11 Euclidean Distance Matrix 511
E.79 Group 12 Euclidean Distance Matrix 512
E.80 Group 13 Euclidean Distance Matrix 513
E.81 Group 14 Euclidean Distance Matrix 514
E.82 Group 15 Euclidean Distance Matrix 515
E.83 Group 16 Euclidean Distance Matrix 516
E.84 Group 17 Euclidean Distance Matrix 517
E.85 Group 18 Euclidean Distance Matrix 518
E.86 Group 19 Euclidean Distance Matrix 519
E.87 Group 20 Euclidean Distance Matrix 520
E.88 Group 21 Euclidean Distance Matrix 521
E.89 Group 22 Euclidean Distance Matrix 522
E.90 Group 23 Euclidean Distance Matrix 523
E.91 Group 24 Euclidean Distance Matrix 524
E.92 Group 25 Euclidean Distance Matrix 525
E.93 Group 26 Euclidean Distance Matrix 526
E.94 Group 27 Euclidean Distance Matrix 527
E.95 Group 28 Euclidean Distance Matrix 528
E.96 Group 29 Euclidean Distance Matrix 529
E.97 Group 30 Euclidean Distance Matrix 530
E.98 Group 31 Euclidean Distance Matrix 531
E.99 Group 32 Euclidean Distance Matrix 532

List of Algorithms

5.1 Module 2: Term Extraction 106
5.2 Module 3: WordNet Relation Extraction 112
5.3 Module 3: Recursive FindRelations SubRoutine 123
5.4 Module 4: Graph Building Basic Algorithm 124
5.5 Module 4: WNDag Graph Construction 128
6.1 Similarity Detection and Difference Identification 150
6.2 Similarity and Difference Graph Building 153
6.3 Measuring similarity and difference 156
6.4 Identifying Key Differences 183

List of Tables

3.1 Meronym Examples 58
3.2 Domain Examples 59
3.3 Derivation Examples 61
3.4 List of unique beginners 63
3.5 List of top hierarchies 64
3.6 Examples of entailment in WordNet 65
3.7 Semantic domains in WordNet for verbs 66

4.1 Contingency table for multiple parentage, χ^2**(8, N=75180) = 324.27, p<0.001** 81
4.2 Hierarchy overlap count 83
4.3 Depth and height values in sub-hierarchies 87
4.4 Polysemy values for nouns, verbs and adjectives 91

5.1 Rank ordering of WordNet relations in (Al-Halimi and Kazman 1998, p.338) 117
5.2 Classification of WordNet relations into directions adapted from (Hirst and St-Onge 1998, p.308) 117
5.3 Fields of WordNet Topological Input File 125
5.4 Sample of WordNet Topological Input File 125
5.5 Sample binary relations file 126

5.6 Categories of Model Information Types 140

6.1 Merged NodeInGraph fields 154
6.2 Features for Weighting Algorithm 157
6.3 Relation Type Weights 165
6.4 Node Specificity Examples 168
6.5 Normalised Node Specificity Examples 168
6.6 System Techniques for Weight Calculation 170
6.7 System Techniques Formulae 171
6.8 Four-way table for similarity measures 179

7.1 List of similarity ratings with associated value 200
7.2 Reduced similarity rating scale with original value mappings . 207
7.3 Reliability Data Matrix of 6 story pairs (6-point similarity scale) 228
7.4 Coincidence Matrix for calculation of Kappa Coefficient 229
7.5 Matrix of Difference Metrics for calculation of Kappa Coefficient 230
7.6 Regression Coefficients 234
7.7 ANOVA, Dependent Variable: Percentage Agreement, Predictors: (Constant) Mean Similarity Rating 235
7.8 Linear Regression Model Summary Table 235
7.9 Descriptive Statistics 235
7.10 Curve Estimate Model Summary Table 236
7.11 Cochran Q statistics for 6 data subsets 240
7.12 Cochran Q statistics for training data 245
7.13 Distribution of Similarity and Relatedness Ratings in Training Data 246

8.1 Threshold values for Human Data 258
8.2 Seventeen System Techniques 264

8.3 *Performance Better Than Chance* Count for Test Case and Similarity Measure Combinations 287
8.4 Successful Results per Technique 288
8.5 Differential Sentences Results for 17 System Techniques 296
8.6 Differential Sentences Results for 8 Story Pairs 298

E.1 Group 1 table of similarity ratings 444
E.2 Group 2 table of similarity ratings 444
E.3 Group 3 table of similarity ratings 445
E.4 Group 4 table of similarity ratings 445
E.5 Group 5 table of similarity ratings 446
E.6 Group 6 table of similarity ratings 446
E.7 Group 7 table of similarity ratings 447
E.8 Group 8 table of similarity ratings 447
E.9 Group 9 table of similarity ratings 448
E.10 Group 10 table of similarity ratings 448
E.11 Group 11 table of similarity ratings 449
E.12 Group 12 table of similarity ratings 449
E.13 Group 13 table of similarity ratings 450
E.14 Group 14 table of similarity ratings 450
E.15 Group 15 table of similarity ratings 451
E.16 Group 16 table of similarity ratings 451
E.17 Group 17 table of similarity ratings 452
E.18 Group 18 table of similarity ratings 452
E.19 Group 19 table of similarity ratings 453
E.20 Group 20 table of similarity ratings 453
E.21 Group 21 table of similarity ratings 454
E.22 Group 22 table of similarity ratings 454

E.23 Group 23 table of similarity ratings 455
E.24 Group 24 table of similarity ratings 455
E.25 Group 25 table of similarity ratings 456
E.26 Group 26 table of similarity ratings 456
E.27 Group 27 table of similarity ratings 457
E.28 Group 28 table of similarity ratings 457
E.29 Group 29 table of similarity ratings 458
E.30 Group 30 table of similarity ratings 458
E.31 Group 31 table of similarity ratings 459
E.32 Group 32 table of similarity ratings 459

Chapter 1

Introduction

In the world today, we are constantly bombarded with information from all kinds of sources. Much of this information is in the written form from warnings to "please mind the step" to theses on Michelangelo, from novels of Russian emigrés to exhortations to "drink Coca Cola". Text comes in an infinite variety of shapes and sizes and styles and colours and yet all text has one key characteristic in common: it all carries meaning. This thesis is inspired by that astounding fact. The main objective throughout the course of this research has been to provide some representation of meaning in text which is empirically-grounded and makes use of resources available but is not subject to them. A second key objective has been to exploit this representation in order to grapple with the vast profusion of text with which we are surrounded. In this research, this secondary objective is formulated as a problem of text differentiation: if one can represent text in a meaningful way, then it should follow that one can differentiate between text content, identifying similarities *and* differences.

The assumptions and approaches taken in any research of this kind must be validated and evaluated rigourously. Therefore, an ancillary focus of this

research has been to ensure that, at all times, the evaluation process is governed by human performance and human judgments. To this end, a key component of the research has been to compile all materials necessary for such an evaluation: corpora of authentic texts and gold standard human judgments of text similarity and difference. The following section sets out how these key objectives were defined as research aims. The final section provides an outline of the rest of the thesis.

1.1 Thesis Aims

The aims of this thesis can be divided into three core areas: computational models of text and their application to the task of text differentiation, analysis of the WordNet knowledge base for use in these models and empirical grounding and validation of the afore-mentioned models.

The three primary computational aims of the thesis are as follows:

1. To design a system to construct a meaningful representation of text within the framework of lexical cohesion that exploits text-based cohesive devices and the semantic content and structural relations of a knowledge base, for the purposes of this research, WordNet. This representation should be derived:

   (a) from free text without mark-up and without disambiguation;

   (b) using all open class parts of speech.

2. To exploit all features of this representation to estimate similarity of content between text documents.

3. To exploit all features of this representation to identify differential information content in text documents.

The computational approach in this thesis employs the WordNet knowledge base as its real world knowledge source and as such could be subject to the vagaries of its hand-coded structure and content. The second set of aims addresses the issue of controlling for the idiosyncrasies of WordNet by means of a detailed analysis of its structure and content. These aims are as follows:

1. To provide a full structural analysis of WordNet in order to estimate the degree of variant data sparseness or density in the knowledge base and to identify the topological features that best characterise its structure.

2. To derive an empirically-grounded metric of concept specificity in the knowledge base which may be used to factor out the non-uniformity of its structure and to evaluate the informational contribution of its individual concepts.

3. To investigate the effects of concept specificity and relation type in a WordNet-based representation of text. This aim is subsumed by the similarity and difference detection goals set out above as concept specificity and relation type are two of the features of the text model which these goals aim to exploit.

The final set of aims addresses the validation of the models and metrics derived according to the afore-mentioned computational and knowledge base analysis goals. These empirical aims are as follows:

1. To compile a corpus of real news texts to be used in the evaluation of a text representation and differentiation model.

2. To derive a gold standard of human similarity judgments to be used in an evaluation of the performance of a similarity estimation algorithm.

The following sections sets out where these goals have been addressed in the thesis.

1.2 Thesis Outline

This thesis describes the empirical and computational approaches that have been taken, during the course of this research, to address the issue of meaningful text representation and comparison within a solid theoretical framework provided by the notion of lexical cohesion in text.

Chapter 2 sets out the theoretical framework for this thesis. The guiding principle of this research from the outset has been meaning in text, how we humans process it and how it may be represented. This chapter outlines the defining characteristics of text as a unified whole, most notably textual coherence and cohesion. Subsequent sections discuss the main approaches to the task of meaningful text representation. These range from unstructured "bag of word" models to highly structured representations governed by the cohesive or coherent structure of text. The research reported in this thesis is conducted within the framework of lexical cohesion. Therefore, the second half of this chapter focuses on the main theories, models and applications of lexical cohesion in the linguistics and computational linguistics domains.

Chapter 3 describes the materials used in the computational and empirical aspects of the research reported in this thesis. The computational approach taken exploits the WordNet knowledge base from the University of Princeton (Fellbaum 1998) as a vital source of real world knowledge. Section 3.2 describes the semantic content, organising principles and some of the inherent weaknesses of WordNet. The empirical studies carried out during the course of this research use three corpora of news texts: the RTÉ, Australian and

Google News corpora. Section 3.3 describes the origin and contents of each corpus.

Any application that relies on a hand-coded knowledge base, such as WordNet, as its source of real world knowledge is subject to the potential difficulties arising from the variant data sparseness of the knowledge base. Chapter 4 aims to address this problem. Section 4.2 presents an analysis of the structure of WordNet, investigating key characteristics of its topology. Section 4.3 describes an empirical experiment carried out to identify which features of WordNet topology are good indicators of concept specificity in the knowledge base. Section 4.4 concludes the chapter with a discussion of how topological features can be used to factor out some of the idiosyncrasies of content and structure in the knowledge base. This idea is expanded in chapters 5 and 6 that follow.

Chapter 5 describes Storylines, a system I designed and implemented to represent text in a meaningful fashion within the framework of lexical cohesive theory. The system exploits text-based cohesive devices, WordNet semantic content and WordNet topological features to construct a meaningful graph representation of text. Section 5.4 to 5.7 discuss the four modules of the program, leading from free text to structured representation. Section 5.8 outlines the data structures used to store the final graph representation. The chapter closes with a discussion of the latent information content of the Storylines text model.

The second primary focus of this thesis is the issue of text differentiation. Indeed, discerning similarities and differences among texts has been a vital concern in the domain of Information Management for many years. Chapter 6 describes the second component of the Storylines system to estimate text similarity and identify differential content between two texts. The chapter

discusses how this component exploits the latent information content of the Storylines text model to accomplish these tasks.

Chapter 7 describes a psycholinguistic experiment to derive a gold standard of human similarity judgments to be used in an evaluation of the Storylines similarity estimation component. Sections 7.3 and 7.4 outline the method and basic results of the experiment. Section 7.5 provides a detailed exposition of methods for assessing reliability of experiment results and their application to this experiment.

Chapter 8 sets out the results of the evaluation process for the Storylines system reported in chapters 5 and 6. The chapter details a thorough evaluation of the similarity estimation component using three gold standard data sets and a pilot experiment to evaluate the differential text content component of the system.

The final chapter, Chapter 9, sets out directions for future work which this research opens up as well as the key contributions of the work and conclusions that arise from it.

Chapter 2

Meaningful Text Representation: Theory and Applications

2.1 Introduction

This thesis aims to provide a meaningful representation of text that may be used in the task of determining semantic similarity of texts. This chapter explores the different approaches to the question: "how to represent text?" The approaches are broadly divided into three categories:

1. Unstructured bag-of-words models;
2. Structured models based on text coherence;
3. Structured models based on text cohesion.

For each category, the main theories are set out, followed by a discussion of the natural language processing problems to which they have been applied.

The chapter opens with a discussion of what is a "text", set out in section 2.2. Section 2.3 discusses models of text which are not based on linguistic theory but rather approach the problem of text representation as one of data modelling. Section 2.4 discusses theoretical models of text representation based on high-level coherence structures and their applications. Section 2.5 discusses the lower level phenomenon of cohesion and theoretical models representing aspects of cohesion. The main focus of this thesis is a particular aspect of this: lexical cohesion. Section 2.6 discusses lexical cohesive devices, structural models of lexical cohesion and implementations of these models and their application to natural language processing tasks.

2.2 Text *vs.* Untext

A fundamental tenet of text linguistics is that a "text" constitutes "any passage, spoken or written, of whatever length, that does form a unified whole"

(Halliday and Hasan 1976, p.1). The difference between text and a collection of sentences effectively considered arbitrarily ordered, an "untext", is patently clear to a reader. The examples of text in contrast with "untext" in (Johnson-Laird 1983, p.356) perfectly illustrate this distinction: one example, the text, makes sense and the other, the "untext" does not, although as a reader one may try to impose sense upon it. This chapter is divided between approaches that aim to model how this distinctive feature of text is realised (sections 2.4, 2.5 and 2.6) and others (section 2.3) which cannot and, indeed, make no attempt to make any distinction between a text and a collection of words or sentences. The latter non-linguistic perspective, in a sense, parallels the human hearer/reader's efforts to interpret "chunks of language which are conventionally presented as texts .. as constituting texts" (Brown and Yule 1983, p.199). However, where humans are concerned with interpretation and meaning, non-linguistic models are only concerned with surface representation.

If we accept this distinction between text and untext, where text is a unified whole and untext is just a sequence of sentences without unity, this begs the question how is textual unity created in a text. De Beaugrande and Dressler (1981) posit seven standards of textuality which all texts must meet, along the lines of the Gricean maxims of conversation (Grice 1975). These standards are intentionality, acceptability, informativity, situationality, intertextuality, cohesion and coherence. The first five standards are concerned with producer and receiver and context of use for the text (for example, a sign on a road as opposed to a poem in a book or an advertisement in a newspaper). Cohesion and coherence, however, are text-based concepts. According to these principles, a text must be both cohesive and coherent, i.e. it must hang together well, and it must make sense. The concept of cohesion

is concerned with linguistic realisation while coherence refers to conceptual structure, as stated explicitly by Johnson-Laird (1983, p.370)[1]:

> a necessary and sufficient condition for discourse to be coherent, as opposed to a random sequence of sentences, is that it is possible to construct a single mental model from it.

Any attempt to model *text* in *this* sense must address the fact that a text comes with a presumption of coherence, fulfilling the requirements of textuality. Without access to the producer, receiver or text context, it is the text-based features of cohesion and/or coherence that will determine that model. Section 2.4 discusses models of text which aim to represent aspects of its coherent conceptual structure and applications of these models. Sections 2.5 and 2.6 discuss theoretical models of text cohesion, with section 2.6 focusing on the primary theoretical framework for this thesis, lexical cohesion. The next section discusses models which do not presuppose any structure in text but rather treat text as a simple sequence of terms.

2.3 Unstructured representations

Successful commercial Information Retrieval, authorship identification and other applications are often based on unstructured representations of text. Such representations treat a document as a bag of words or characters (as opposed to a set as here multiple instances are recorded). The basic premise is that words or characters alone provide a suitable and sufficient representation with which to characterise document content. Furthermore, word order

[1]Johnson-Laird's (1983) hypothesis of a "**single** mental model" is not a comment on processing ambiguities in natural language. The mental models theory allows for ambiguity in representations.

is not relevant, beyond the value of n chosen for n-gram models where $n > 1$. Such a representation does not make use of linguistic knowledge as there is no attempt to derive structure from or impose structure on the text. This kind of approach has the advantage that the computational overhead of building a text representation is low. In linguistic terms, such a text model is theory-independent and, for this reason, I will discuss it in terms of applications rather than underlying theory. A fundamental difference between different "bag of word" models lies in what ends up in the bag. Some models use characters rather than words. However, in most cases, a set of terms is derived after initial pre-processes of stopword removal[2] and stemming[3] are undertaken. All "bag of word" rather than "bag of character" models are based on the resulting set of stemmed document terms. The models, however, differ in what information is encoded with the document terms. This can range from a Boolean field indicating presence or absence in a document to extensive corpus-based collocational information.[4] These basic models have been applied to a vast range of Natural Language Processing and speech tasks, from word sense disambiguation (Yarowsky 1995) to text segmenta-

[2]The process of removing those words from the bag which are deemed information-poor. These usually include closed class words (e.g. determiners, prepositions), frequently occurring words (e.g. verbs "be", "have") or semantically "vague" terms (e.g. "something", "anywhere").

[3]The process of identifying the root form of words in a document. This step entails some morphological analysis of terms. The level of analysis can range from removing plural and tense markers to more complex derivational morphology.

[4]These models constitute one end of a broad spectrum of approaches that represent text using the words or lexis of a text alone. Lexical cohesive models belong to the structured end of this spectrum where the words are assumed to form particular structures in text. The research reported in this thesis can be included at that end of the spectrum where the words are used as the key to unlocking textual structure and content.

tion (Choi 2000), authorship identification (Medori 2004), learning analogies (Turney and Littman 2003) and mapping document collections (Smith 2000), as well as information retrieval (Baeza-Yates and Ribeiro-Neto 1999). The following sections outline the details of three "bag of word" models and enumerate some related applications.

2.3.1 Term Frequency Vector Models

At its most simple, a "bag of word" model may be just a list of document terms. Most commercial IR systems, however, use a weighted term vector space model where term weights are based on term frequency. The vector is such that it, typically, contains a frequency count for all terms, and is thus indexed by terms. These are also applicable to character models of text. In a simple term frequency model, weight might be equal to the frequency of a term in a given text or frequency normalised for document length. More successful models, however, take account of local (document) and global (collection) information. Salton and McGill's (1983) classic Vector Model calculates term weight w_i according to the following formula:

$$IDF = log\frac{D}{df_i} \tag{2.1}$$

$$w_i = tf_i * IDF \tag{2.2}$$

The tf_i term is the number of times term i occurs in a document (usually normalised for document length). IDF, the inverse document frequency, is calculated according to formula 2.1 where D is the number of documents in the collection and df_i the number of documents containing term i. This model captures something of the distribution of individual terms in a particular document collection. Very common terms will have low weights and uncommon terms will be assigned high weights when they appear often in

a text. This approach models language according to a Zipfian distribution where a small minority of words appears very frequently but most other terms are extremely infrequent in language (Zipf 1935). Some researchers have posited the need to redefine this basic distributional model in order to take account of the "clumpiness" of words (the fact that there is a significantly greater probability of encountering a word in a text once it has been encountered already in that text). Church and Gale (1995) and Katz (1996) factor this into their models of term frequency in corpora.

The fact that this model is entirely word-based in its published applications constitutes both a key advantage and a significant drawback of the approach. It is an advantage in terms of processing but a drawback from the point of view both of data sparsity and synonymy. The problem of data sparsity is particularly significant for short texts. The problem of synonymy is a more profound difficulty, however. Identical content could potentially be realised lexically in several distinct ways. While the vector space of a document aims to model the complete vocabulary environment of a text, the resulting model can easily be biased by a particular author's lexical preferences. Imagine two texts about the confectionary item, *KitKat*, where one describes it as a biscuit and the other as a bar, the model would not necessarily relate the texts. This issue could prove to be a drawback for information retrieval but an advantage to applications such as authorship identification where it is nuances of surface structure, not content, that are of interest.

2.3.2 Explicitly Stochastic models

A second approach to unstructured text representation uses statistical language modelling techniques. This approach has been central in speech recognition primarily, where language models are used to predict the probability

of sequences of phonemes or bi-phones or words belonging to the language modelled (Bahl *et al.* 1989, Kuhn and Mori 1990). This premise translates well to text-based applications where sequences of character or word n-grams are predicted. This model assumes that in order to predict a word w, only the previous $n-1$ words are relevant, as in equation 2.3, where i is the position of the word in the text indexed from the start and where n gives the size of the window of text used to make an estimation.

$$P(w_i|w_1...w_{i-1}) = P(w_i|w_{i-n+1}..w_{i-1}) \tag{2.3}$$

The weights are calculated on a training corpus. Computational limitations mean that the size of n must be limited to the range 1 – 6.[5] Due to the Zipfian distribution of language noted above, one of the central open areas in language modelling research remains how to deal with zero frequency, i.e. encountering words or n-grams never encountered during training. Various smoothing techniques have been suggested and it is often this smoothing that determines the success of a given language model. Aside from speech applications, language modelling has been applied to several text-based processing tasks such as part-of-speech tagging (Brants 2000), information retrieval (Ponte and Croft 1998, Lafferty and Zhai 2001, Hiemstra 2001) and text categorization (Peng and Schuurmans 2003).

2.3.3 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is also a statistical model of word usage (Deerwester *et al.* 1990). It is, however, motivated primarily by research in text comprehension. Its primary assumption is that "there is some underlying or "latent" structure in the pattern of word usage across documents"

[5]Indeed, the unigram model, where $n = 1$ and probabilities are calculated on single words rather than word sequences is a feasible and popular method.

(Foltz 1996, p,198). This approach is close to lexical cohesive models on the "bag of word" — structured models continuum. However, it is discussed here with other unstructured text representations as LSA does not define what *kind* of structure it aims to extrapolate from text. Essentially, it is a model of word collocations in document collections. Unlike the term vector models which treat words as static features of text, this model is concerned with word "meaning", where the "meaning" of a word is approximated by its relations to other words. While term vector models count term occurrences, LSA counts term co-occurrences generating a similarity matrix for all terms in a document. Using Singular-Value Decomposition (SVD), the similarity matrices are reduced to some number k orthogonal factors. SVD, like Principal Components and Factor Analysis, is a mathematical tool to extrapolate underlying factors that best explain patterns, in this case, patterns of vocabulary co-occurrence, in data. Using this technique, words with similar patterns of co-occurrence will have similar vectors in the reduced LSA factor space. The assumption is that words with a similar meaning will have similar patterns of co-occurrence. Hence, this model is ideally suited for evaluating similarity among texts. Each text is represented as values in the k dimensional space derived from training co-occurrence matrices and document similarity is calculated as the dot product or cosine of their vectors in this space.[6]

The derivation of a model using Latent Semantic Analysis is very sensitive to several aspects of the training phase. Firstly, the model assumes a degree of homogeneity between the training data and the data which the sys-

[6]The distance between documents is conceptualised as the size of the angle between their representative vectors. Hence, the cosine of the vectors is used as a measure of similarity.

tem is intended to model. Distributions of words and word collocations will be very different across text types.[7] Therefore, the factors produced for one training corpus may not be extendable to another. Secondly, the selection of the dimensionality, k, for the subspace representation can have an enormous impact on the derived representation. Foltz (1996) set the optimum number at $100 < k < 300$. Within this broad range, however, the results for document similarity judgments can vary widely (Lee *et al.* 2004). Finally, the co-occurrence counts may be weighted, similar to the term vectors discussed above and this choice can greatly affect performance of the trained system also. Despite these possible pitfalls, LSA has been used to estimate document similarity for a number of different purposes (measuring text coherence (Foltz *et al.* 1998), analysing student essays (Wolfe *et al.* 1998, Weimar-Hastings *et al.* 1999), text segmentation (Choi *et al.* 2001), problem solving (Quesada *et al.* 2002) and text categorization (Laham 1997)).

2.4 Coherence: Structured Representations

Although useful, the models discussed in the previous section, from a linguistic point of view, are not designed to model "text" but rather collections of words or word sequences. This approach is satisfactory from a computational (and indeed commercial) perspective but is not adequate for modelling the coherent structures that are texts. Researchers from several intellectual sub-fields have addressed this problem of defining discourse structure . If we accept that discourse or text is a *structured* collection of segments, researchers must investigate what determines this high-level structure and how it is con-

[7] See (Kilgarriff 2001) for a discussion of the difficulties of determining corpus homogeneity.

structed in text. The main approach is to view coherent text as a series of related segments and subsegments so that each segment of text is connected to its adjacent segment by some coherent relation. In the same way as syntactic theory addresses a clause as a structured hierarchy of sub-clausal units, so discourse semantics views discourse as a hierarchical structure where all sub-segments are connected into one unifying whole. Theories differ widely, however, in the number, organization and nature of the intersegment relations they claim provide structure and coherence in discourse. The following sections outline two central theories of textual coherence, Rhetorical Structure Theory § 2.4.1 and Grosz and Sidner's (1986) intention-based approach, § 2.4.2.[8] Section 2.4.3 discusses attempts to synthesise the two theories. Section 2.4.4 goes on to outline some applications of these theories to Natural Language Understanding and Generation tasks.

2.4.1 Rhetorical Structure Theory

One of the most popular discourse theories of the last 15 years is Mann and Thompson's (1988) Rhetorical Structure Theory (RST). This theory is not psychologically motivated but is driven largely by research in the area of text generation. The central notion of RST is the rhetorical relation, a relation which holds between two non-overlapping text segments called nucleus and satellite. This nucleus-satellite distinction reflects the assumption that

[8]This account ignores many other analyses of coherence in text. Notable among the exceptions are the formal approaches of Hobbs (Hobbs 1990, Hobbs 1995), Polanyi (1988) and Asher and Lascarides's (1994) "Structured" extension of DRT (Kamp and Reyle 1993). The exceptions are made, firstly, because the two theories discussed here constitute the two major poles of debate in the area in the last fifteen years and, secondly, because coherence is not the main focus of this thesis although aspects of the relationship between coherence and cohesion are discussed in sections 2.5.3 and 2.6.2.

some text segments are more important than others in a discourse. They posit that a nucleus is "more essential to the writer's purpose" (Mann and Thompson 1988, p.266), and that omission of the nucleus would destroy text coherence, whereas omission of a satellite would not. They set out a taxonomy[9] of 23 relations, the relevant constraints that must be satisfied for each and the effects each produces. The list of relations may be subdivided into those that relate to the informational structure of the text (e.g. elaboration, cause, circumstance) and those that relate to its intentional structure (e.g. motivation, enablement, justify). According to this theory, the coherent structure of a text may be represented by a single Rhetorical Structure Tree composed of a hierarchical organisation of sub-trees representing relations between text spans. The theory imposes a uniqueness constraint on relations, i.e. only one relation may hold between adjacent text spans.

A clear benefit of this theory is its formalisation of constraints and effects for a finite set of rhetorical relations which provide an analysis of coherence in text. For this reason, it has proved popular in computational models of discourse (see section 2.4.4). However, criticism has been levelled at the basic theory for various reasons. The first and perhaps primary criticism is that it does not fully specify the intentional structure of discourse, what the writer wants to achieve with the text. It might appear that this is due primarily to the uniqueness constraint, whereby only a single relation may hold between spans of text. However, Moore and Pollack (1992, p.543) would argue that, in fact, informational and intentional structures are not even necessarily isomorphic and, hence, it is a non-trivial task to determine a mapping between these levels in RST. A second problem is the issue of how the relations are

[9]Although the relations are generally presented as list, they are often grouped in subgroups which would suggest a taxonomic organisation.

realized in text. Mann and Thompson (1988, p.244) state that the theory describes relations "whether or not they are grammatically or lexically signalled". Extensive work has been carried out on mapping RST relations to cue phrases (Marcu 1997b, Knott and Mellish 1996).[10] However, according to Harabagiu (1999), only 25% of sentences contain cue phrases to indicate their coherent structure. Therefore, the work of Matthiessen and Thompson (1988), Bateman and Matthiessen (1993) and others on the lexicogrammatical realization of RST relations should be expanded to investigate the full range of features that realise RST relations in text and by which RST relations can be identified.

2.4.2 Grosz and Sidner (G&S)

A second core theory which is often contrasted with RST is Grosz and Sidner's (1986) theory based on intentionality in discourse. In this case, the structure of a discourse is determined by the intentions utterances are meant to convey, the Discourse Segment Purpose in their terminology. The basic premise, then, is that coherent discourse is possible through hearers' understanding of speakers' intentions. That is, coherence is the result of hearer's assimilating what a speaker wants to achieve with a text. While they do not reject the notion of rhetorical relations, they argue, in contrast with RST, that it would be impossible to define a finite set of relations that hold between discourse segments. Instead, two basic intention-based structural relations are posited, *dominance* and *satisfaction-precedence*. The derived structure for discourse is, again, a tree with a single over-arching Discourse Purpose, which is composed of sub-trees defined by the two afore-mentioned relations. The theory also addresses the attentional structure of discourse

[10]See section 2.5.3 for a brief discussion of cue phrases.

with a dynamic model of objects, properties and relations that are salient as the discourse progresses.

The broad scope of the theory, encompassing several of De Beaugrande and Dressler's (1981) standards of textuality, make this a very attractive theory. The focus on intentions in discourse has meant that it is particularly popular in the field of Natural Language Generation (NLG). However, the theory hinges on determining discourse boundaries and the primary intentions of the speaker at each segment but does not give clear criteria for tackling these tasks, thereby limiting its application in Natural Language Processing tasks. The difficulty of assigning segment boundaries has been the subject of much research.[11] The question of extrapolating intentions from text is more difficult and is all but left to the intuitions of a discourse analyst. For this reason, this approach, on its own, is more suited to text analysis than to computational applications other than NLG. However, as noted above, a theory of discourse is incomplete without some account of intentionality such as this theory provides.

2.4.3 Bridging the Gap between RST and G&S

The previous two sections outlined two core theories of discourse structure. Debate between exponents of the two approaches has been rife over the last decade as it is clear that both have much to offer across several disciplines, from discourse analysis to text generation, and yet both have some drawbacks. Some researchers, notably (Moore and Pollack 1992, Moser and Moore 1995, Hovy 1991, Marcu 2000), have looked at synthesising aspects of the two. Hovy (1991) suggested including a results field in RST rhetorical relations to add an intentional level to informational relations or an informa-

[11] See Passonneau and Litman (1997) for a review of studies in the area.

tional level to intentional relations. Moser and Moore (1995) set out a similar approach, Relational Discourse Analysis (RDA), which takes its intention-based focus from Grosz and Sidner's (1986) theory and its pair-wise relations from RST. The authors note that the notion of nuclearity in RST parallels the dominance relation in Grosz and Sidner's theory. This is translated into the *core:contributor* relation in RDA. RDA, however, relaxes the uniqueness constraint of RST. Therefore, each contributor is linked by an informational *and* an intentional relation to a core. Marcu (2000) takes the work of Moore and Moser as starting point and formalizes a unified theory for RST and G&S. Within this framework, intentions are used to reduce possible RST interpretations while RST structures are used to identify intentions in the discourse.

There remains the problem of the non-isomorphic nature of information and intentional structures. While it is true that several different RST trees may be derived from the same text, to my knowledge, it is not necessarily the case that any of these trees will coincide with an intentional analysis of that text. Moore and Pollack (1992), for example, describe the structural mismatch in example 2.4 between an intentional analysis that identifies coming home by 5 as the nucleus of the discourse and an informational analysis wherein finishing the bookshelves is the nuclear proposition.

(2.4) Come home by 5:00. Then we can go to the hardware store before it closes. That way we can finish the bookshelves tonight (Moore and Pollack 1992, p.544).

Only Marcu (2000) addresses this issue, conceding that non-isomorphic structures could not be maintained simultaneously in his model. He notes, however, that his model will enumerate all possible rhetorical structures and their associated intentions and allow researchers to study the relationships

between the intentional and informational structure and determine if and when non-isomorphism arises.

2.4.4 Applications of Theories of Coherence

The high-level structure of textual coherence defined by the theories outlined above can and does play a crucial role in many natural language applications. Natural Language Generation and particularly text-planning applications are often designed to exploit the coherent structure derivable from rhetorical relations (McKeown 1985, Hovy 1993, Knott 1991, Moore and Paris 1989, Rosner and Stede 1992). Indeed, Moser and Moore's (1995) hybrid theory derives from observations that RST trees lacked adequate intentional structure for text planning in natural language generation (Moore and Paris 1992). A second area that benefits from representations of the coherent structure of text is summarization (Marcu 1999, Power 2000, Bouayad-Agha *et al.* 2000). The nuclearity principle of RST, in particular, provides a vital indication of which elements of text can be omitted and which are central to the meaning of a text. The need for a unified account of discourse in terms of information and intention has also been addressed in this area (Rino and Scott 1996). Coherence structures have also been used to improve the output of machine translation systems (Al-Onaizan *et al.* 2000, Tucker *et al.* 1986) and to evaluate essays automatically and provide students with feedback (Miltsakaki and Kukich 2004, Burstein and Marcu 2000).

2.5 Theory of Cohesion: Structured Representations

Coherence is concerned with the high-level informational, intentional and attentional structure of text and how this is integrated with a receiver's model of the real world. Cohesion, on the other hand, is a grass-roots characteristic of text concerning how individual pieces of text hang together. Much current research is indebted to Halliday and Hasan (1976) for their exposition of cohesion as a combination of forces at work to knit a collection of disparate sentences into a text, a cohesive whole. A text is defined and sustained by cohesive ties, relations of meaning, that are set up between elements of the text. More succinctly:

> Cohesion occurs where the INTERPRETATION of some element in the discourse is dependent on that of another. The one PRESUPPOSES the other, in the sense that it cannot be effectively decoded except by recourse to it. (Halliday and Hasan 1976, p.4)

In example 2.5, the possessive pronoun *her* refers back to *Sofia* and *they* in the second sentence refers back to *her eyes* in the first. It is these anaphoric ties which mean that we, as readers, interpret the sentences as a *text*, rather than as two separate sentences.

(2.5) Well, Sofia has been opening her eyes a lot. Every day they seem more brown.

In this section, I discuss "the set of semantic resources for linking a sentence with what has gone before" (Halliday and Hasan 1976, p.10) which Halliday and Hasan categorize into five devices that contribute to establishing *texture*

in text. These devices are the three related cohesive ties of reference, substitution and ellipsis (§ 2.5.1 and § 2.5.2), conjunction (§ 2.5.3) and lexical cohesion (§ 2.6).[12]

It is worth, at this point, noting a distinction that Brown and Yule (1983) make forcibly in their re-exposition of Hallidayan cohesion:

> An important distinction needs to be drawn ... between the "meaning relations" which hold between items in a text and the explicit expression of those "meaning relations" within a text. (Brown and Yule 1983, p.195)

While the discussion of cohesive devices below is made in terms of explicit markers, it does not make the claim that these *markers* are necessary in order for readers to accept a text as text. It is rather the underlying semantic relations that these markers realise that are essential to creating text. Conversely, use of these markers does not guarantee that a collection of sentences is accepted as text, unless they are used to realise genuine semantic relations.[13] With this caveat in mind, there follows a discussion of the four cohesive devices of reference, substitution, ellipsis and conjunction and how they contribute to cohesive text. As the primary focus of this thesis, the discussion of lexical cohesion, its structures and applications, is set out in greater depth in section 2.6.

[12]It is outside the scope of this thesis to provide a formal or computational account (syntactic or semantic) of reference resolution, ellipsis or conjunction. The role of this review is to elaborate recent thought on how the concepts discussed contribute to an "omniscient" theoretical perspective of texts. My system, discussed in chapters 5 and 6 are not omniscient but are motivated by this theoretical framework.

[13]See (Enkvist 1978, p.110) for an example of cohesive markers used in an incoherent non-text.

2.5.1 Reference

The cohesive device of reference comprises those elements in a text which cannot be interpreted on their own but rather with reference to something else either in the text or within the context of situation of the text. This device then includes exophora (referent outside the text), anaphora (referent precedes the referring expression) and cataphora (referent follows the referring expression). Example 2.5 illustrates two examples of anaphora. Examples 2.6 and 2.7 are instances of exophoric and cataphoric reference respectively.

(2.6) Isn't she beautiful! (talking about a baby in the context of situation)

(2.7) They are great fun these holidays

According to Halliday and Hasan (1976), this relation is a *semantic* one as it is not subject to grammatical constraints. The referring expression, whether endophoric or exophoric, can be of any grammatical class, regardless of the referent. This distinction is relevant in delineating reference from substitution which does impose the identical grammatical class requirement. Furthermore, the phenomenon of reference is not restricted to pronominal reference but includes demonstratives, the definite article, comparatives and demonstrative adverbs as in examples 2.9 to 2.12.

(2.8) Sofia was born in New York on the evening of the 29th of September.

(2.9) *That* evening was one to remember.

(2.10) *The* evening brought us such delight.

(2.11) The *same* evening we toasted her with champagne.

(2.12) I can't wait to go *there* and see her.

2.5.2 Substitution and Ellipsis

Where reference can set up cohesive ties between nouns and possessive determiners or locations and demonstrative adverbs, substitution is classified as those cases where an element of a particular grammatical category is substituted with one of the same category. This formal relation is realised for nominals by substitution with "one" and "same", (2.13 and 2.14), for verbs with "do" (2.15) and for clauses with "so" (2.16).

(2.13) I dropped my phone down the toilet. I'll have to get a new *one* soon.

(2.14) J.R. ordered a scotch on the rocks, so I got the *same.*

(2.15) Do you want a biscuit? No, but she *does.*

(2.16) It's nearly finished. Or, at least, I hope *so*!

Ellipsis is a special case of substitution where the element in the text is substituted for by nothing. For example, nominal substitution replaces a head noun with "one", nominal ellipsis removes the head noun altogether. Examples 2.17 to 2.19 illustrate the use of ellipsis as a cohesive tie.

(2.17) Nominal: He just got me a pint of Guinness. Where's yours?

(2.18) Verbal: Could she be finished by now? I don't think she could.

(2.19) Clausal: Why did you leave me there? I didn't.

As grammatical constructs, these devices are subject to many restrictions. Constituents may not be omitted or substituted at will in the production of well-formed cohesive text.

2.5.3 Conjunction

Conjunction is a semantic relation which like the grammatical relations discussed above serves to connect a text element with what went before. The nature of the connection, however, is very different from the aforementioned three: the cohesive tie in this case specifies how the following segment is to be interpreted relative to what went before. Examples 2.20 to 2.23 below illustrate this point.

(2.20) They sped around town in the sports car. She went into labour at 2 in the morning.

(2.21) They sped around town in the sports car. *As a result*, she went into labour at 2 in the morning.

(2.22) They sped around town in the sports car. *Subsequently*, she went into labour at 2 in the morning.

(2.23) They sped around town in the sports car. *In spite of that*, she went into labour at 2 in the morning.

The conjunctions in italics each create a distinct type of cohesive tie between the first and second sentence. The juxtaposition of the two propositions in the first example might suggest some interpretation but the inclusion of an explicit conjunction creates a strong cohesive tie that directs the reader in their interpretation. Halliday and Hasan (1976, pp. 242-243) set out a detailed taxonomy of conjunction types and usage. These conjunctions are a point of overlap in the cohesive and coherent structure of text. They are both overtly cohesive and intrinsically rhetorical. While reference, substitution and ellipsis may not impact on the high-level coherent structure of discourse, these items are the cue phrases which determine local elements

of that structure. The system I developed effectively ignores the textual resources represented by conjunction. However, an intended extension of the work is to exploit such rhetorical markers, as well as logical functions, such as negation, to provide a richer structure that is more representative of the *coherent* structure of text while maintaining the rich cohesive dimension (see section 9.2).

2.6 Lexical cohesion

The previous section outlined four of the five cohesive devices posited by Halliday and Hasan (1976). These constitute key resources in building cohesive structure, each of which is restricted by certain formal criteria for use. This section discusses the final aspect of cohesion, lexical cohesion, which is not subject to formal constraints and could be said to be largely responsible for the richness of associations in cohesive structures. The examples of grammatical cohesion in sections 2.5.1 and 2.5.2 may seem somewhat wooden and this is due to the fact that they do not exploit the potential of this key cohesive device of association between lexical items.[14] In practice, lexical cohesion is just part of the spectrum that defines how words are associated with each other in text. At one end of the spectrum, there is reference and substitution, which define how elements of the discourse may be referred to by semantically empty, or nearly empty, lexical items, such as pronouns and "one" and "does". At the other end is lexical cohesion, which is concerned with how semantically rich lexical items are used to create associations whether through co-reference or collocation. The selection of varied lexical items carrying their

[14]Indeed, it is a somewhat contrived task to single out one type of cohesive device when, usually, in text all devices are interleaved to build a rich cohesive structure.

own lexical content and with their own associations to refer to and/or replace elements in the text is an extremely prevalent device and one whose analysis and application has proved very popular in the linguistics and computational linguistics domains.

The excerpt below is drawn arbitrarily from Wilde's (1908) *The Picture of Dorian Grey.* In this sentence alone, a rich set of associations are woven from the lexical items. There are the threads of perfume, of flowers and of a dwelling. This tapestry is generated not through an interplay of grammatical relations but from the juxtaposition of lexical items.

> The studio was filled with the rich odour of roses, and when the light summer wind stirred amidst the trees of the garden there came through the open door the heavy scent of the lilac, or the more delicate perfume of the pink-flowering thorn. (Wilde 1908, p.1)

Lexical cohesion provides a powerful means of establishing threads that can run through a whole text, holding it together as threads in a tapestry. Admittedly, these threads or chains tend to be more banal, less brightly coloured and scented in everyday texts than in this excerpt but they constitute the most promiscuous cohesive ties in text. What makes the notion of lexical cohesion so interesting is that it is not restricted to the narrow confines of co-reference. Indeed, "properly speaking, reference is irrelevant to lexical cohesion" (Halliday and Hasan 1976, p.284). It is concerned with the patterns of lexical choices which a producer of text makes, firstly, to create a cohesive text and, sometimes, secondly, to locate that text within a genre, text type or topic area. This section sets out the different ways of achieving lexical cohesion and discusses how the lexical patterns within texts, and between texts, are structures worthy of study in their own right.

2.6.1 Lexical Cohesive Devices

Reiteration

Reiteration is the form of lexical cohesion that regards repeated reference to entities in text. It covers the myriad of different ways of referring to the same entity in a text. The spectrum of possibilities runs from simple repetition and general nouns, which lie in the grey area between substitution and reiteration, through synonyms and on to all kinds of elaborate paraphrase using subordinate or superordinate terms. Often the new lexical item is accompanied by a referring expression such as "the", "this" or others. Example 2.24 sets out a list of noun phrases all of which refer to the entity "my new car".

(2.24) my new car: the machine — the vehicle — the car — the automobile — that pink Fiat — the car being towed away by the clampers

These techniques of reiteration are roughly paralleled by Hoey's (1991) definitions of simple and complex repetition and simple paraphrase.

Collocation

The collocates of a word are those words with which the original word habitually keeps company and with which they are associated.[15] Halliday and Hasan (1976, p.284) state that this type of cohesion is "achieved through the association of lexical items that regularly co-occur". Hoey's (1991) account leans not towards co-occurrence but rather towards relatedness of meaning. Of complex paraphrase, which may be roughly equated with Halliday and Hasan's (1976) definition of collocation, Hoey (1991, p.64) states that this

[15]Philips (1985) notes the words are often inter-correlated such that collocates of one term will often also be collocates of the other.

occurs "when two lexical items are definable such that one of the items includes the other". Whatever the definition, collocation is the relation that holds between non-coreferential expressions which are conceptually related. This includes systematic relations between agents, patients and instruments of the same theme, antonyms, meronyms, holonyms and other non-systematic associations that hold between entities in the real world or between expressions in text.[16]

The short news text in example 2.25, from the RTÉ Corpus, illustrates the use of collocation to achieve cohesion in text. The list of terms related to pregnancy and birth, set out in example 2.26 threads its way from beginning to end of the text. Indeed, the humour in the opening line of the expression, "an unexpected delivery", relies on how the term "delivery" ties in with other lexical, birth-related items in the text, overriding its more common meaning.

(2.25) Gardai in Dublin had an unexpected delivery last night, when a heavily pregnant woman gave birth while en-route to the city's Coombe Hospital. The woman stopped off at Rathcoole on her way to the hospital and requested a Garda escort to get her to the labour ward on time. The baby was born in the Garda station. (RTÉ News Interactive, 4/12/2002)

(2.26) delivery — pregnant — birth — Coombe Hospital — labour ward — baby — born

The notion of collocation is where lexical cohesion can expand beyond the borders of individual texts. Hoey, in an address to the IVACS–04 conference

[16]This last catch-all relation type is motivated by the need to highlight associations humans make between expressions due to their co-occurrence, if only once, in seminal texts of their culture. For example, Italians tend to associate the word *selva*, forest, with *oscura*, dark, as this collocation is prominent in the opening lines of Dante's *Inferno*.

in Belfast, expounded how the filaments of collocation relations function intertextually as well as intratextually. In an in-depth analysis of several texts on the subject of the discovery of a planet beyond Pluto, some unusual collocates in these texts were found to derive from original comments made by the discoverer of the "mysterious planet X". Beyond the domain of discourse analysis, it is the task of literary criticism to tease out these intertextual collocational threads in order to determine the effects and associations that a writer may be trying to produce or reproduce.

2.6.2 Structure: Nets and Chains

Lexical cohesive devices do not create relations simply between pairs of words. As example 2.26 above suggests, lexical items form chains of related items through text. Morris and Hirst (1991) coin the phrase lexical chain for these sequences of related words. A text can and often does contain several chains of related items of varying length which through their interaction make text cohesive. In text 2.25 above, besides the "birth" chain, there is also a Garda chain (2.27).

(2.27) Gardai — Garda escort — Garda station

Together, these chains carry much of the specific meaning of the text and contribute to its cohesive structure. Hoey (1991) further develops this idea as repetition nets. There is a shift in focus here from the chains of lexical items to the sentences linked by these chains. An in-depth analysis of parallels between pairs of sentences linked by multiple chains highlights the interactions between these chains and how they contribute to the coherent structure of text. The processes by which the parallels are derived are lexical, syntactic and discoursal, covering the range of processes readers use to interpret text.

This extra dimension of chain interaction is essential if we are to consider the full contribution of lexical patterns to text *meaning* as a whole. Indeed, Hoey's (2004) more challenging claim is that the phenomenon that we call lexical cohesion within text, and which corpus linguists investigate as collocation across texts, is dictated and defined by "lexical priming", the notion that lexical items prime particular other lexical items. Lexical priming implies a new theory of the lexicon, and language in general, which is driven by the intertextual associations of lexical items that have long been understood as a fundamental component of what it means to *know* a word.

The potential to expand the role of lexical cohesion to more structural aspects of text requires more insight into the interactions within the patterns created by reiteration and collocation. From a computational perspective, however, the chains alone provide an appealing means of access to text content which by-passes other elements of text structure. Section 2.6.3 sets out how this has been achieved in particular implementations for deriving cohesive structure. Section 2.6.4 goes on to discuss how these structures, and other unstructured approaches based on lexical cohesion, have been exploited in natural language applications.

2.6.3 Lexical Chaining Algorithms

If we accept that lexical items combine to form cohesive chains through text, a key problem for a computational implementation of cohesion remains how to effectively and efficiently derive the "correct" chains from text automatically. This section sets out Morris and Hirst's (1991) original chaining algorithm and variations to their technique by subsequent researchers as well as Harabagiu and Moldovan's (1998) related search engine approach.

Morris and Hirst

Morris and Hirst's (1991) chaining algorithm defined a method for generating lexical chains from text using Roget's Thesaurus (Roget 1977). The algorithm was not implemented at the time but has been the basis of most lexical chaining systems since. Candidate terms for chaining were selected from all non closed-class words. The stopword removal process also excludes high frequency words as these were deemed too information-poor to be of use in the lexical chain representation. The first candidate term in the text becomes the head of the first lexical chain. All subsequent terms either become members of an existing chain or become the head of a new chain. Membership of a chain is evaluated on the basis of three criteria:

1. Thesaural relations: Morris and Hirst (1991) identify five types of relations that hold between words and their indices in the thesaurus. 90% of all chain links were based on two of these:

    - Two words share a common index category (e.g. residentialness — apartment)
    - An index category of one word points to an index category of another (e.g. car — driving)

2. Transitivity of relations: some relations are deemed stronger than others as they are less likely to introduce spurious chains.[17]

3. Distance between words in a chain: there is a limit to the distance allowable between chain elements. In this algorithm, the limitation

[17]See section 3.2.2 for a discussion of WordNet relations and chapter 4 for a discussion of features of the topology of WordNet that may be exploited to interpret these relations.

was quite restrictive.[18] Subsequent implementations have relaxed this constraint.

The outcome of this algorithm is a series of chains of related words. Morris and Hirst used these chains to identify intentional segment boundaries in text as an initial step in determining the under-lying intentional structure of the text as defined in the theory of Grosz and Sidner (1986).

Hirst and St-Onge

The Morris and Hirst algorithm was the basis for many subsequently implemented chaining algorithms. It was first implemented by Okumura and Honda (1994) and later on Roget's Thesaurus by Ellman (2000). Most other implementations have used the WordNet knowledge base, not only because it is free and relatively up-to-date, but also it has a well-defined structure and a library of access functions. Hirst and St-Onge (1998) provided the first such implementation, applied to the task of detecting malapropisms in text. They are responsible for mapping Morris and Hirst's (1991) thesaural relations to equivalent relations in WordNet. These are divided into three categories:

1. Extra strong: repetition;
2. Strong: synonymy, hypernymy—hyponymy, holonymy—meronymy, antonymy;
3. Medium: any *allowable* path between WordNet concepts, or in WordNet terminology synsets, up to a maximum path length of 5.[19]

[18]The word distance restriction was in fact a design feature of the Morris and Hirst (1991) algorithm as the aim was to identify intentional segment boundaries in text.

[19]See section 3.2 for a full exposition of WordNet relations and in particular section 3.2.1 for an explanation of the WordNet synsets.

The concept of *allowable* paths is aimed to limit spurious connections between synsets and is based on direction type and allowable changes of direction of links in the WordNet hierarchy. Like Morris and Hirst (1991), the candidate words for chaining are selected from all open class, high frequency terms in the input text. Hirst and St-Onge, however, limit these terms to all nominals. They state that this is due to the lack of connectivity between part-of-speech hierarchies in WordNet.[20] The algorithm first attempts to chain words within their local context of the sentence. It then checks the previous context to identify possible chain relations, if this fails it introduces a new chain headword. Many other chaining algorithms (Green 1999, Kozima and Ito 1997, Stokes 2004) have been based on this system and have maintained their restriction to exclusively nominal chains.

Barzilay and Elhadad

Barzilay and Elhadad (1997) developed a version of this algorithm, whereby all possible chains are generated and the best, those which are the most densely connected and have the highest weights, are selected at the end of processing. The invocation of the selection process only after the generation of all possible chains represents a non-greedy approach to the task of lexical chaining building. Their candidate term selection process also differs from that of Hirst and St-Onge. The text is first tagged. All noun phrases are extracted and compound nouns are reduced to their head noun (e.g. "quantum computing" to "computing"). Hirst and St-Onge do not tag the text and rely on the chaining algorithm itself to disambiguate between noun and verb

[20]The system I developed which is reported in chapter 5 extends the conventional focus of many lexical chaining systems to date to other syntactic categories, namely verbs and adjectives. The extra connectivity of WordNet version 2 makes this extension feasible.

senses of words. Barzilay and Elhadad (1997) impose a further constraint on lexical chain construction that chains may not cross text unit boundaries. These units are determined using Hearst's (1997) Text-Tiling algorithm.[21]

Stairmand

Stairmand's (1996) algorithm is also a non-greedy approach to lexical chaining using WordNet. The set of candidate terms again is the set of nouns in the text. All possible links between each sense of each word in the text is computed. These links are then clustered and clusters with shared senses are merged. This merging corresponds to the Hirst and St-Onge step of finding allowable paths between words in WordNet. Wherever points in a cluster exceed the allowable inter-chain element distance of 80 words, the clusters are broken into smaller chains. Chain weights are assigned according to text span covered by a chain and the number of terms in the chain (chain density).

Harabagiu and Moldovan

Harabagiu and Moldovan (1998) implemented a somewhat different approach to lexical chain building. The system uses a knowledge base derived from WordNet version 1.5. incorporating information from the WordNet glosses. The pre-processing assumed here greatly exceeds that required for the Morris and Hirst (1991) algorithm: all candidate words are assumed to have been disambiguated and thematic roles assigned. This system takes all parts of speech as input as the extended knowledge base provides the necessary inter-part-of-speech connectivity. The lexical chaining component is conceptualised as a search engine to find semantic connections between knowledge

[21] See section 2.6.4 for a brief discussion of Hearst's (1997) and other text segmentation applications.

base concepts. The search originates from the two concepts which should be connected and is successful when the two search chains meet up at a common concept. The propagation relations for simple connections through the hierarchy for nouns are hypernymy and meronymy, for verbs hypernymy, entailment and cause, for adjectives similar and pertainym and adverbs only the pertainym connection. The system also searches gloss connections — those connections derived from connections between tagged role-assigned elements of the WordNet glosses. The system searches simple and gloss connections in parallel. The resulting paths are then filtered based on relevance as well as reiteration, density and path length, as in Morris and Hirst (1991).

One might expect that this algorithm would provide a much richer structure as it is composed of all parts of speech and uses concept definitions in the WordNet glosses. Harabagiu (1999), however, compares the output of their system with that of Morris and Hirst (1991). Despite the different assumptions, both return similar paths, although Harabagiu's algorithm identifies more pairs of words and so has higher recall in evaluations. Perhaps due to the demanding pre-processing assumptions and to the use of the derived extended WordNet knowledge base, while this algorithm is used in many applications in Harabagiu and Moldovan's home institution of the University of Texas, it has not been re-applied elsewhere.

2.6.4 Implementations of Cohesion

The previous section discussed algorithms for deriving lexical chains from text. This section sets out some of the main natural language applications to which the notion of lexical cohesion in general has been applied. Some applications use lexical chains to achieve their ends, others are based on a less structured notion of lexical cohesion. The range of applications is vast and

shows how the lexicon alone provides a means of approaching most natural language problems.[22]

Discourse Analysis

Morris and Hirst's (1991) original algorithm was aimed at extrapolating discourse structure from text, based on patterns of lexical cohesion. They found that "the lexical chains ... correspond closely to the intentional structure produced from the structural analysis method of (Grosz and Sidner 1986)" (Morris and Hirst 1991, p.40). Harabagiu (1999) expanded on this basic concept in a system that builds lexical chains from texts, automatically deduces coherence relations and builds a coherence structure for a given text. The coherence structure built by Haraagiu's system is induced from lexical paths, cue phrases and the pragmatic knowledge encoded in the WordNet knowledge base, modelling "the correlation between textual cohesion, discourse cue phrases and textual coherence" (Harabagiu 1999, p.248). The motivation for this approach is that only 25% of sentences contain discourse cues (Harabagiu 1999), therefore, coherence structure must be based on more than cues alone. This system deals with three categories of relation covering the top level of Hovy's hierarchy (Hovy 1993): resemblance, cause-effect and contiguity. In a comparison with Marcu's Rhetorical Parser (Marcu 1997a), the results of an experiment to derive the rhetorical structure of forty Wall Street Journal articles were almost 80% identical. As Marcu's (1997a) system is based entirely on discourse cues, this might lead one to question the utility of the lexical cohesion component of his system. The representations built, however, have been applied to the problem of text inferencing with

[22]The list of tasks in this section is not exhaustive but they are the most common applications of lexical cohesion in the literature.

some success (Harabagiu and Moldovan 1998).

Other algorithms for text segmentation and topic detection are based on Morris and Hirst's (1991) assumption that

> Lexical chains tend to delineate portions of text that have a strong unity of meaning. (Morris and Hirst 1991, p.23)

Hearst's (1997) TextTiling algorithm exploits the cohesive device of repetition to determine segment boundaries in text. Kaufmann's (1999) VecTile system works on a similar premise to LSA (see section 2.3.3) whereby word similarity is measured as the cosine between word collocation vectors. VecTile uses two adjacent 10-word windows to build up a similarity curve based on these vectors where troughs represent low cohesion, indicating a topic change, i.e. a section break. A comparison of the performance of these two segmentation systems would suggest that, while collocation (represented by VecTile) is a key factor in cohesion, under-emphasis of the significance of simple repetition can adversely effect performance on a segmentation task. These systems lie on the boundary between "bag-of-word" models and systems that fully exploit the potential of lexical cohesion in text.

Lexical chains have also been applied to the related task of topic detection and tracking in the on-line broadcast news domain (Carthy and Smeaton 2000, Stokes and Carthy 2001, Carthy 2002, Stokes *et al.* 2002, Stokes 2004). The cohesive structure built by Stokes's (2004) LexNews algorithm uses WordNet thesaural relations augmented with news domain-specific statistical word associations and an analysis of proper noun repetition to detect new topics and, thereby, segment an input news stream.

Word Sense Disambiguation and Semantic Similarity

There are two main approaches to word sense disambiguation (WSD) deriving from the notion of lexical cohesion. The first approach, taken by several researchers, applies measures of semantic similarity or relatedness between words connected by lexical cohesive links to the task. The second approach views word sense disambiguation a by-product of the lexical chaining process, as in chain construction some senses of polysemous words will be discarded.

The semantic similarity approach was implemented initially by Sussna (1993) as an edge counting technique based on relative depth of concepts in a knowledge base. Leacock and Chodorow (1998) took a similar approach based on short paths in a knowledge base. The basic assumption to this approach is that the word sense with the shortest path connections to other words in the text is probably the relevant word sense for that text. Such approaches, of course, suffer from the inaccuracies and uneven distribution of most knowledge bases. Subsequent research has focused on developing a more accurate metric of semantic similarity or relatedness. Agirre and Rigau (1995) introduced the notion of conceptual density as a measure of similarity. Conceptual density is sensitive to the following four measures:

1. shortest path length between concepts;
2. depth in the hierarchy — the deeper in the hierarchy, the closer the concepts;
3. concept density within the hierarchy — concepts in a dense part of the hierarchy are relatively closer than in a sparse area;
4. independence from the number of concepts being measured;

Equation 2.28 sets out the formula, discussed in detail by Agirre and Rigau

(1995).

$$CD(c,m) = \frac{\sum_{m-1}^{i=0} nhyp^i}{descendants_c} \tag{2.28}$$

The numerator expresses the expected area of a subhierarchy of some concept c containing m marks or senses of the words to be disambiguated. The divisor is the actual area of the subhierarchy. The formula, then, gives the ratio of weighted marks under c to the number of descendants of c, i.e. the relation between the weighted marks and the total area of the subhierarchy. $Nhyp$ represents the mean number of hyponyms per node. This is calculated to represent the relationship between the height, average number of hyponyms and total number of senses in the subhierarchy, if it were homogeneously distributed. The notion of conceptual density and semantic relatedness is relevant to this research as the concept weighting algorithm discussed in chapter 6 and the concept specificity experiment discussed in section 4.3 are designed to identify and use a similar measure of semantic comparability that is motivated by human responses. Resnik (1995) developed a hybrid metric of similarity which was also applied to the task of semantic and syntactic disambiguation (Resnik 1999). This hybrid approach uses a measure of information content, from information theory (Ross 1976), derived from corpora in the estimation of relatedness among concepts. Other semantic similarity approaches have exploited other aspects of the source text, the knowledge base and corpora (Rosso *et al.* 2003, Mihalcea and Moldovan 2000, Fernández-Amorós *et al.* 2001, Harabagiu and Maiorano 1999).[23]

The lexical chaining by-product approach to word sense disambiguation has also been applied often in the literature. Indeed, the first implementation of the (Morris and Hirst 1991) chaining algorithm was tested on a word sense disambiguation task for Japanese by Okumura and Honda (1994).

[23]See Budanitsky (1999) for an extensive survey of measures of semantic relatedness.

Subsequent systems, which have often been largely based on Hirst and St-Onge's (1998) algorithm, have proved quite successful at word sense disambiguation in view of the difficulty of the task (Stairmand 1997, Galley and McKeown 2003, Stokes 2004).[24]

Summarization

Lexical cohesion and lexical chains have also been used in the construction of summaries from input documents. Barzilay and Elhadad's (1997) chaining algorithm constituted the analysis component of a summarization system. The heuristics for extracting key sentences for the summary were identified by lexical chain criteria such as text units containing the first appearance of a chain head word or high concentration of chain constituents. Silber and McCoy (2002) developed a linear-time version of the Barzilay and Eldahad algorithm which included a chain evaluation component. The system was evaluated on a small corpus of scientific articles and textbook texts and their abstracts. The intuition here is that the main concepts are represented by lexical chains that have been evaluated as strong and therefore these should be the chains present in the abstract. The results report precision and recall of almost 85% on the identification of strong chains in source and summarized text. However, it is difficult to interpret this result as the authors themselves note that there is no baseline result for comparison, nor is it obvious what kind of baseline to employ. Several other systems have since developed this approach for single and multiple document summarization (Brunn *et al.* 2001, Chali *et al.* 2003, Alemany and Fuentes 2003). Moving away from lexical chaining, other systems have used lexical cohesion in the construction of summaries. For example, Mani and Bloedorn (1999) used

[24]Stokes (2004) reports precision and recall of 59% and 56% respectively.

cohesive links derived from WordNet and phrase relations in the task of multi–document summarization and Jing (2000) uses cohesive links in text to determine focus in the local context and, thereby, identify which elements may be removed during summarization.

Text Classification and Semantic Orientation

Classification techniques based on lexical cohesion are based on the premise that a representation of cohesive ties in a text is an appropriate and adequate representation of its semantic content. Ellman (2000) used Morris and Hirst's (1991) chaining algorithm for Roget's Thesaurus (Roget 1977) to derive a "Generic Document Profile" for texts. This profile is a set of Roget's categories with associated weights calculated on the basis of chain length and density. This generic profile for individual text was used to compare texts using a case-based reasoning approach. The system was evaluated on encyclopaedia texts with human similarity judgments as the gold standard.[25] It was found that in three out of six subject areas, the results of Ellman's system were significantly correlated with the human results. Green (1999) takes a similar approach in the task of building hypertext links between documents. This system determines inter-document similarity as the cosine of vectors of WordNet synsets for each document. Using Hirst and St-Onge's (1998) chaining algorithm, document synset vectors are constructed that consist of all the synsets in all of the lexical chains identified in a document or document segment. The system was evaluated by an analysis of links followed by a group of subjects on a collection of newspaper articles. Green found that subjects did not note any advantage in using the hyperlinks generated by the

[25] Chapter 7 sets out the experiment I designed to derive human similarity judgments that were used as a gold standard in the evaluation of the system I developed.

chaining system over links generated by a term vector space model.

An area of application related to text classification involves determining semantic bias or orientation of text within the context of lexical cohesion. In this instance, the positive or negative bias of lexical items is estimated to determine the overall orientation of a text. Researchers have addressed this issue by evaluating the semantic orientation of terms in a knowledge base, such as WordNet, and compiling an overall metric of semantic orientation of the text as a whole based on the collection of lexical items in the text (Wiebe *et al.* 1999, Hatzivassiloglou and Wiebe 2000, Turney 2002, Wiebe *et al.* 2003, Kamps *et al.* 2004). This approach takes an unstructured view of cohesion in text.

2.7 Discussion

In this chapter, I have explored the different approaches to representing text both conceptually and computationally. Section 2.2 discussed what it means to be a "text" and the motivations for viewing text as more than a simple "bag-of-words". Section 2.3 outlined the main "bag of word" models of text representation. The term vector space and language models in particular have proved very successful in commercial natural language applications. These kind of models, however, ignore those dimensions of text by which humans interpret text as a unified and meaningful whole. Sections 2.4 and 2.5 explored two of these dimensions: coherence and cohesion. Section 2.4 presented two opposing theories of textual coherence: Rhetorical Structure Theory (RST) and Grosz and Sidner's (1986) intention-based theory (G&S). Both theories define coherent structure as a hierarchical tree structure of relations between non-overlapping discourse segments. RST provides a clear

taxonomy of the informational and intentional relations by which this structure is organised. G&S define how intentions and attention dictate construction and interpretation of the structure. The central aspects of both theories are essential to a unified theory of discourse coherence and section 2.4.3 outlined some attempts to provide this unified account. While the structures defined by these theories are essential in a comprehensive representation of text content, in practice, it is a non-trivial task to derive the structures from text either automatically or by hand. Despite this inherent difficulty, theories of coherence have provided the theoretical framework for many natural language generation and processing applications (section 2.4.4).

Sections 2.5 and 2.6 set out Halliday and Hasan's (1976) theory of textual cohesion, the factors that make text hang together well. Section 2.6 deals with lexical cohesion in particular. This section sets out the kind of relations that lexical items in text enter into and sections 2.6.3 and 2.6.4 elaborate how these have been exploited in many areas of natural language processing research in recent years. Within the framework of lexical cohesion, a key premise is that the semantic content and structure of text can be accessed through the lexical items in a text. Such a model is not dependent on other aspects of linguistic knowledge, such as syntax or rhetorical structure or segment boundaries, for its functionality, although it can, of course, benefit and be enriched by such information.

The research reported in this thesis is firmly rooted in this approach to text representation and to solving problems in natural language processing. The aim has been to take plain text as input and exploit the lexis of that text, not just nouns but verbs and adjectives also, in order to construct a meaningful representation of the source text. The approach taken here deviates from Halliday and Hasan's (1976) lexical chain structures which have been

employed widely in the NLP domain (see section 2.6.4). Rather it is inspired by the network structure described in (Hoey 1991) (see section 2.6.2). The key insight here is that lexical chains do not exist in isolation but interact with each other throughout the text. In this thesis, text is conceptualised as a network or graph structure where elements of different grammatical categories inter-connect and these interconnections, as well as the elements themselves, carry meaning. The system I developed to build this network structure from free text exploits the WordNet knowledge base in order to derive these underlying semantic relations between words in text. The precise contribution of the knowledge base is controlled through an analysis of its organising principles (section 3.2) and topology (chapter 4). Details of the implementation of this graph representation are set out in chapter 5. Chapter 6 discusses how this representation has been applied in the task of estimating text similarity and difference.

Chapter 3

Materials: Knowledge Base and Corpora

3.1 Introduction

The previous chapter set out the theoretical framework for this thesis. This chapter describes the materials used in the computational and empirical aspects of the research reported in subsequent chapters. The computational approach taken exploits the WordNet knowledge base from the University of Princeton (Fellbaum 1998) as a vital source of real world knowledge. Section 3.2 describes the semantic content, organising principles and some of the inherent weaknesses of WordNet. This section is intended as an introduction to the knowledge base for those unfamiliar with the system. Chapter 4 provides the novel contribution of a full structural analysis of WordNet which guides the interpretation of the knowledge base in the Storylines implementation described in chapters 5 and 6. The empirical studies carried out during the course of this research employ three corpora of news texts: the RTÉ, Australian and Google News corpora. Section 3.3 describes the origin, contents and, where appropriate, the compilation process for each corpus.

3.2 WordNet

Many theories and applications in computational linguistics assume or require some real-world knowledge source as an integral part of a working system. For many, this knowledge source is provided by WordNet. According to the originators at the Cognitive Science Laboratory in Princeton, WordNet is an on-line lexical database (Fellbaum 1998). This section examines the knowledge base to illustrate that it is more than just a machine-readable dictionary and its functionality extends beyond mere enhanced dictionary look-up. While the dictionary-lookup aspect of the knowledge base is fast and hassle-free, its potential and the ambitions of the designers are far greater as

their account of the motivations for developing the database illustrate. The database was designed with the intention of providing a model of the mental lexicon, such as has emerged from psycholinguistic or more specifically psycholexicology research over the last fifty years. As such, the macro and micro-structure of WordNet are motivated by theoretical properties of the mental lexicon and not historical evidence of a language, as in a traditional dictionaries such as the Oxford English Dictionary.

The research documented in this thesis exploits the functionality of WordNet in order to represent text meaning. This section sets out the main organizational principles of WordNet which are assumed in subsequent chapters. Section 3.2.1 discusses the Synset, the main unit of meaning in WordNet consisting of sets of synonymous words, and the originators' definition of synonymy as the key organizing principle in the database. Section 3.2.2 discusses the relational rather than compositional nature of semantics in WordNet. Section 3.2.3 outlines how nouns are organised in inheritance hierarchies in the database and sets out the central concepts at the head of these hierarchies. Sections 3.2.3 discusses troponymy, the organising principle of verbs in WordNet and how the definition of the verb structure differs from the noun hierarchies in the knowledge base. Section 3.2.3 outlines how modifiers are represented in their relationship to nouns in the knowledge base or independently as clusters of modifiers of similar semantic content. The design and implementation of WordNet is well documented in (Miller and Fellbaum 1991, Fellbaum 1998), therefore, the following sections aim only to set out the key concepts, not to document the system exhaustively.

3.2.1 Synonymy and the Synset

A key innovation of WordNet is what it defines as the main unit of the database. Where the main unit of a traditional dictionary is the *word form*, the main unit of WordNet is the *word meaning*. It is an accepted feature of most languages that many word forms have more than one associated meaning and often belong to more than one syntactic category. The lexical form "article" is one such example of a noun which has several independent associated meanings. In a dictionary, this lexical form has one entry, listing several different senses of the word form "article". In WordNet, however, the word form "article" appears in four independent entries, one for each different meaning of the lexical form, as in examples 3.1 to 3.4.

(3.1) article – (non fictional prose forming an independent part of a publication)

(3.2) article – (one of a class of artefacts; "an article of clothing")

(3.3) article, clause – (a separate section of a legal document (as a statute or contract or will))

(3.4) article – ((grammar) a determiner that may indicate the specificity of reference of a noun phrase)

In this sense, the database is structured like a thesaurus rather than a dictionary. The functionality of the database, however is designed such that, regardless of its under-lying structure, WordNet may function as a dictionary or as a thesaurus according to a user's needs.

The principle, then, by which the vast array of lexical information of a language[1] is organised is **synonymy**. Every unit in WordNet represents a

[1]WordNet was developed for English but WordNets for many other languages are currently under development. EuroWordNet is perhaps the most complete non-English

single meaning and this meaning may be lexicalised by several different synonymous word forms. In WordNet terminology, this unit is called a **Synset** or a set of synonymous words. As the key organizing principle, synonymy must be clearly defined by the designers of the database. Synonymy of words, in the strict sense, where two word forms are *always* interchangeable without altering meaning of a statement, is extremely rare. The definition of synonymy in WordNet is a more relaxed concept which facilitates the grouping of "synonymous" words into Synsets. Synonymy in WordNet is, in fact, synonymy in context, which is defined as follows:[2]

> two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value. (Fellbaum 1998, p.6)

By this definition, synonymy is reciprocal, i.e. if word a is synonymous with word b, word b is synonymous with word a. This reciprocal definition of synonymy in context is the key assumption for concept definition within WordNet. In all, the knowledge base contains 152,059 word forms organised into 115,424 Synsets, groups of synonymous word forms, which correspond to all the meanings the word forms express. Word forms, then, often appear in many independent Synsets, as in the "article" example above.

A Synset represents a lexicalised concept, that is a concept which is represented by some set of lexical forms. These lexical forms are not restricted to single word forms but include multiword expressions such as "jack of all

database and has been implemented in Dutch, Italian, Spanish, German, French, Czech and Estonian with an Inter-Lingual Index which provides the mapping between concepts in the individual EuroWordNets and Princeton WordNet.

[2]Although synonymy is defined in truth conditional terms, in practice, it is viewed as the top end of a semantic similarity gradient.

trades" or "Deus ex machina", and phrasal verbs such as "bring up" or "look forward". A concept may be lexicalised by several different such lexical forms. As regards a formal definition of concepts, WordNet sidesteps the question of what, in fact, is the concept that word forms lexicalise. In their introductory discussion, Fellbaum (1998) note that :

> These synonym sets (synsets) do not explain what the concepts are; they merely signify that the concepts exist. (Fellbaum 1998, p.5)

From the perspective of the psychological literature on concepts, this is an unsatisfactory definition as, in WordNet, a concept is merely defined by the word forms that express it and this definition entails the assumption that to know a concept, one must first know the meaning of the word forms that express it. From the perspective of lexicology, this is a reflection of the eternal conundrum of what to use to express word meaning besides words themselves. These issues are not solved nor are they fully addressed in WordNet.

Synset contents

This section addresses the composition of synsets in the WordNet framework, leaving aside the fact that concepts are not formally identified in theory, nor operationally, within the knowledge base. The primary content of a synset is a set of synonymous word forms. For computational reasons, each synset is assigned a unique eight digit numeric ID, such as those in example 3.5, termed a synset offset, which is the byte offset of the synset in the data storage file.

(3.5) 00270602, 00323358, 01615477, 00004358, 00004483

The part of speech for each synset is recorded as a single character.[3] The number of synonymous word forms in the set is logged. Furthermore, each synset contains a count of the number of synsets to which this synset is related and pointers of different kinds to these related synsets.[4] The final component of the synset is a gloss of the concept, similar to a dictionary definition and, often, some examples from corpora of the word forms in use.

The gloss contains much information, such as the distinctive features of concepts, which might be usefully encoded in a different format in the knowledge base. Harabagiu *et al.* (1999) discuss an extended WordNet with extra relations derived from these glosses. The Extended WordNet project, on-going at the University of Texas in Dallas, aims to provide an enhanced knowledge source using gloss and corpus-derived information. To date, the potential of the synset glosses is not fully exploited in the WordNet database.

Examples 3.6 and 3.7 set out the synsets for "zebra" and "snore", respectively. The "zebra" synset, from the noun data file, contains only one word form. The "snore" verb synset consists of a set of three synonymous words.[5]

(3.6) 02303885 05 n 01 zebra 0 005 @ 02287933 n 0000 #m 02287627 n 0000 ~02304070 n 0000 ~02304209 n 0000 ~02304344 n 0000 |any of several fleet black-and-white striped African equines

(3.7) 00016718 29 v 03 snore 0 saw_wood 0 saw_logs 0 006 * 00014429 v 0000 @ 00001740 v 0000 + 00785431 n 0101 + 06941747 n 0101 + 09925437 n 0101 + 00785431 n 0102 01 + 02 00 |breathe noisily during one's sleep; "she complained that her husband snores"

[3]Section 3.2.3 outlines the division of WordNet and its constituent data files into four syntactic categories.

[4]Section 3.2.2 discusses the types and semantics of the relations defined in WordNet.

[5]The explanation of all components of a synset entry are documented in the WordNet on-line documentation at http://www.cogsci.princeton.edu/ wn/doc.shtml.

The first 8-digit number is the synset offset. The gloss is the text following the vertical bar. All other 8-digit numbers are related synsets and the symbols @, ~, #m, and + represent the kind of relation that holds between the source and related synsets. These relations will be discussed in detail in section 3.2.2.

3.2.2 Relational semantics

The organizing principle of WordNet is relational. The semantics of a concept are not defined in terms of semantic components but rather are inferred by the relations into which they enter with other concepts. In WordNet, these relations can be seen as links among synsets. Therefore, the semantic content of a synset is inferred from the word forms that express it and the kinds of relations it enters into with specific other synsets. These relations are:

- synonymy and antonymy;
- the IS-A hypernymy—hyponymy relations;
- the part-whole meronymy-holonymy relations;
- inter-syntactic category relations, such as domain membership and word form derivation relations.

These relations are categorised as semantic (word meaning *to* word meaning) or lexical (word form *to* word form). All are reciprocal, that is, if x is related to y, then y is related to x. In some case, the relation is symmetrical and, hence, the reciprocal relation has the same name, e.g. antonymy. In other cases, it is not and the reciprocal relation has a different name, e.g. hypernymy and hyponymy.

Synonymy and Antonymy

Section 3.2.1 discussed the role of the synonymy in defining the main unit of WordNet, the Synset. Synonymy in WordNet is defined as interchangeability of word forms in context. It is symmetric, that is, all members of a synset are deemed synonymous with all other members of that synset. There is no symbol to represent synonymy as all synonymous terms are in the same synset.

Antonymy is the relation that holds between a word form x and another word form y that denotes *not* x, as illustrated by the following examples:

(3.8) good and evil

(3.9) fast and slow

(3.10) succeed and fail

Antonymy is represented in WordNet data files with the symbol "!". This relation is lexical, that is, it holds between word forms rather than word meanings. For example, "succeed" is an antonym of "fail". The synonyms of "succeed" (win, come through, etc.), however, are not. Antonymy is the organizing principle for modifiers (adjectives and adverbs) in WordNet (see section 3.2.3). In fact, this brief outline of antonymy belies the complex nature of this lexical relation. However, a more profound discussion is beyond the scope of this thesis and I refer the reader to the documentation in (Fellbaum 1998) for a more in-depth analysis.

Hyponymy / Hypernymy: @ and ~

The hypernymy-hyponymy relation pair defines a semantic relation that holds between synsets or word *meanings*. This relation is often termed the

IS_A relation. It holds between synset A and synset B where "A is a B", or A is accepted by native speakers to be a kind of B. In this definition, the superordinate concept (B) is a hypernym and the subordinate concept (A) its hyponym. Hyponymy allows concepts to be organised in an inheritance system (Touretsky 1986) whereby subordinates inherit all the features of their superordinates and have some added features of their own. For example, a flamingo is a bird. It inherits all of the features of birds: it lays eggs, it has feathers and wings. It also has some distinctive features of its own: it is pink, has long legs and lives in lakes. This relation is the organizing principle for nouns in WordNet (see section 3.2.3).

Meronymy / holonymy: # and %

Meronymy and holonymy define the reciprocal part-whole relation that holds between concepts such as "finger" and "hand, "minister" and "cabinet". More explicitly, a meronym x is accepted by native speakers of a language as being a part of some holonym y. In the examples above, "finger" and "minister" are meronyms of "hand" and "cabinet" respectively. Winston *et al.* (1987) enumerate six ways in which this meronymy relation is realised. WordNet implements the following three, examples of which are set out in table 3.1:

1. the relation between a whole and the members that compose it;
2. the relation between a whole and the substance of which it is made;
3. the relation between a whole and its constituent parts.

Meronymy interacts with the inheritance system set up by the hyponymy relation in an interesting fashion. Meronyms can be seen as distinguishing features of concepts within the WordNet methodology. As such, they are inherited by the subordinate concepts, or hyponyms, of the whole to which

Relation type	Part	Whole	WN Symbols
Member	cabinet minister	cabinet	#m %m
Member	Defence Secretary	US Cabinet	#m %m
Substance	eiderdown	duvet	#s %s
Substance	tissue	organism	#s %s
Part	eiderdown	eider duck	#p %p
Part	lesson	education	#p %p

Table 3.1: Meronym Examples

they are linked, their holonym. For example, table 3.1 states that tissue is a substance of which an organism is made. Therefore, all hyponyms of organism, from a heteroploid to fireman, are also made of tissue. Likewise, a cabinet has cabinet ministers for members. The hyponyms of cabinet, British Cabinet and US Cabinet, are also composed of member ministers. It follows that the superordinates of a meronym are also meronyms of the holonym and its subordinates. For example, organisms are made of tissue which is a body part. Therefore, organisms are made of body parts. It does not follow that the subordinates of a meronym are also meronyms of a holonym and its subordinates. For example, an organism is not necessarily made of cartilage, which is a subordinate of tissue. This relation is only employed in the noun hierarchy. A part-whole relation is meaningless for modifiers and verbs.

Cross-syntactic category relations

The relations discussed above, in WordNet, are relations that hold between synsets of the same syntactic category. Other relations, both lexical and

semantic, are defined that cross syntactic category boundaries. These are a relatively new phenomenon in WordNet. Indeed, a major criticism of early releases of the database was the lack of connectivity between categories. In its current implementation, the knowledge base still does not encode links to indicate selectional restrictions or noun-verb links to identify noun function. The relations that do exist indicate derivational connections between word forms, or domains of usage of word forms. The need to improve semantic connectivity across the database is widely acknowledged and is within the remit of the eXtended WordNet project (see section 3.2.1).

Domains WordNet 2 introduced a new kind of relation to classify synsets according to their domain of use. Three kinds of domains are specified: geographic region, category and usage. Table 3.2 sets out some examples of domain topics and domain terms for each domain type. There are 594 synsets

Domain type	**Domain**	**Domain Terms**
Geographic Region	Ireland	banshee, ceilidh
	Vienna	Wiener schnitzel
Category	Cookery	to season, to puree
	Law	bail out, or 587 other terms
Usage	disparagement	tree-hugger, effeminate
	poetry	darkling, alliterate

Table 3.2: Domain Examples

designated as topical domains. Of these, 422 synsets are category topics, 160 region and 30 usage.[6] These 594 topics classify 8,063 synsets (5,962 accord-

[6]Some synsets are topics of more than one type of domain, hence, the total of 612 synsets.

ing to category, 1,226 region and 981 usage). In total, approximately 7% of the database is coded for domain of use. The relation has been coded as a semantic relation between synsets or word meanings rather than as a lexical relation between word forms or between a synset and a word form. It is not apparent whether the use of a synset to synset relation is theoretically founded or merely a restriction imposed by development resources. Regardless of this possible short-coming, the domain pointers constitute an important advance in cross-category semantic connectivity in WordNet.

Attribute The term attribute is perhaps misleading for this noun-adjective relation. The link does not denote that the adjective is a distinctive feature or attribute of the noun, as in the noun-adjective pairs below.

(3.11) canary - yellow

(3.12) garlic - pungent

Instead, the attribute relation holds between a noun which denotes some attribute and the adjective or adjectives that express values of this attribute. Examples of the WordNet attribute link are set out in the list that follows.

(3.13) reassurance - reassuring, worrisome, comforting

(3.14) gender - male, female, androgynous

(3.15) nature - good-natured, ill-natured

In this way, 337 noun synsets are directly linked to 630 adjective synsets. Given the organisation of adjectives in WordNet which will be outlined in section 3.2.3, this relation links the 337 nouns indirectly to substantially more adjectives in the database.

Derivation The remaining cross-category links are lexical relations regarding derivation of word forms. For nouns and verbs, the relation indicates that word forms are derivationally related (it does not indicate which form derived from which). Nouns and adjectives are linked by the pertainym relation, denoting that an adjective pertains to a given noun. Examples 4 and 5 in table 3.3 illustrate that although generally the lexical form of the adjective is derived from the related noun, this is not always the case. Adjectives are also linked to verbs, where any participle commonly used as an adjective is linked to the verb from which it derives. The final derivational link to mention is between adjectives and adverbs where adverbs are linked to the adjectives from which they derive. Table 3.3 sets out some examples for each of the connections outlined above.

Syntactic categories	**Examples**
Noun-Verb	confusion - confuse
	hypnosis - hypnotise
Noun-Adj	academia - academic
	tooth - dental
	dentistry - dental
Verb - Adj	babble - babbling
	hydrogenate - hydrogenated
Adj - Adv	vast - vastly
	exclusive - exclusively

Table 3.3: Derivation Examples

3.2.3 Syntactic categories in WordNet

A key difference between WordNet and a traditional dictionary is that WordNet is divided into four separate sections according to part of speech. Furthermore, it does not contain function words as these are assumed to be stored separately from the rest of the mental lexicon (Garrett 1982). The division of the lexical database according to part of speech is implemented for two practical reasons:[7]

1. the synonymy principle: synonymous words are interchangeable in context, according to the WordNet definition of synonymy, and therefore must be of the same syntactic category

2. organizational principles: each part of speech is organised according to *different*, empirically justified, criteria. The principles organizing nouns cannot be applied successfully to verbs, adjectives or adverbs and *vice versa.*

The following sections outline the basic structure of the database for each of the four syntactic categories: nouns, verbs, adjectives and adverbs.

Nouns in WordNet

Nouns constitute by far the largest component of WordNet. There are 114,648 nominal word forms in WordNet 2 which are allocated among 79,689 synsets. These synsets are organised as an inheritance hierarchy according to the hyponym relation. As outlined in section 3.2.2, this structure allows for inheritance of features from superordinates to subordinates and avoids redundancy of information in the lexicon — distinctive features are inher-

[7]For psycholinguistic justifications see (Fellbaum 1998, 2).

ited rather than duplicated.[8] The concept or concepts at the top of the inheritance hierarchies are of primary importance in defining the ontology of concepts in the knowledge base. The designers of WordNet defined 25 concepts that they term *semantic primes* or *unique beginners* each of which is the head of a hierarchy which are initially separate but where there may be some overlap of shared concepts. The 25 unique beginners are listed in table 3.4. These concepts were deemed sufficiently generic and yet distinct

act, action, activity	feeling, emotion	plant, flora
animal, fauna	food	possession
artefact	group, collection	process
attribute, property	location, place	quantity, amount
body, corpus	motive	relation
cognition, knowledge	natural object	shape
communication	natural phenomenon	state, condition
event, happening	person, human being	substance
		time

Table 3.4: List of unique beginners

enough to provide a place for all nouns of English without too much over–lap between the hierarchies. These primes are regrouped under 9 top hierarchies which are listed in table 3.5. The selection of these unique beginners was carried out in a systematic fashion by Philip Johnson-Laird. However, it is possible to argue that a different selection could have been made. This point is made merely to highlight the fact that, although WordNet is a very useful tool, users must be wary not to make the assumption that it is a final and

[8]See chapter 4 for an in-depth analysis of the shape of the hierarchy.

Entity	State
Abstraction	Phenomenon
Group	Event
Act, action, activity	Possession
Psychological feature	

Table 3.5: List of top hierarchies

exhaustive representation of concepts in English.

Verbs in WordNet

While nouns are organised according to the hyponymy relation, the overarching principle for verbs in WordNet is entailment. In fact, entailment is encoded in the verb structure by means of four relations. The first such relation is troponymy, the verb version of the hyponymy relation.[9] This links verbs where verb x entails verb y and the two verbs are co-extensive in time. This entailment could perhaps be expressed as verb x is a way of doing verb y. The first column of table 3.6 provides examples of verbs related by troponymy. The second entailment relation is termed entailment, where the verbs are not temporally co-extensive but the action of verb x occurs within the same time period as the action of verb y. The final two verbal entailment relations are causation and antonymy.[10] These relations, as they are encoded in WordNet, generate a structure which is partially an inheritance structure

[9]This term was coined by Fellbaum (1998, p.47) to express this verbal version of hyponymy. The symbol for the hyponym pointer for nouns and the troponym pointer for nouns are the same.

[10]See (Fellbaum 1998) for details.

through the troponym relation. The other entailment relations, however, do not imply inheritance of distinctive features. The verb inheritance hierarchy is much more shallow than the noun hierarchy. Indeed, it is better visualised as a series of clusters rather than a hierarchy with more specific synsets clustering around their superordinate synsets with further connectivity via entailment, causation and antonym relations.

Troponymy	**Entailment**	**Cause**
kiss - touch	arrive at - move	mobilize - move
hurt - upset	snore - sleep	hurt - suffer
gloss - shine	gloss - rub	remind - remember
feed - provide	dub - translate	feed - eat

Table 3.6: Examples of entailment in WordNet

Verbs were divided into fifteen semantic classifications by a similar process to nominal semantic prime selection process. The classification is set out in table 3.7. Fourteen of the verbal semantic primes denote events or actions in specific semantic domains. The stative verb group comprises verbs which denote states rather than actions. This group tends to be organised in a similar fashion to modifiers in the knowledge base, in terms of synonymy and antonymy rather than entailment and causation. The details of this structure as it is realised for adjectives is outlined in the section that follows.

Modifiers in WordNet

The adjectives included in WordNet can be divided into two groups: descriptive adjectives and relational adjectives, which do not describe a property of

bodily care and functions	consumption	perception
change	contact	possession
cognition	creation	social interaction
communication	emotion	weather
competition	motion	stative verbs

Table 3.7: Semantic domains in WordNet for verbs

the referent but rather modify its reference. In WordNet, these two groups are encoded in two distinct ways.

Relational Adjectives The Pertainym relation discussed in section 3.2.2 is used to associate relational adjectives with the noun to which they pertain. They number approximately 3,500 adjectives. They do not enter into other relations with other synsets in WordNet and as such constitute a unique and separate component of the database.

Descriptive Adjectives The remaining 15,000 adjectives can be termed descriptive adjectives. The key relations in the organisation of these adjectives in WordNet are opposition (the *antonym* relation) and similarity (the *similar to* relation). The structure is defined as follows:

> clusters of adjectives associated by semantic similarity to a focal adjective that relates the cluster to a contrasting cluster at the opposite pole of the attribute. (Fellbaum 1998, p.28)

A proto-typical example is the good-bad opposition. The adjectives good and bad are opposing poles of the goodness/badness attribute. Around good,

there is a large cluster of similar adjectives denoting a range of goodness from fair, through good to excellent. Likewise, around bad there is a large cluster of similar adjectives denoting different degrees of badness from poor through bad to awful (and beyond). The connection to the adjective at the centre of the cluster is encoded with the *similar to* relation. This relation allows the lexical relation of antonymy to act as the backbone of the semantic clustering model. The lexical forms at antonymic poles stretch their tendrils out to all similar adjectival synsets to form semantic clusters. This structure has been used in applications to detect semantic orientation of adjectives and text as outlined in section 2.6.4.

Adverbs Adverbs in WordNet are not organised according to a central principle. The adverb component of the database comprises approximately 4,650 word forms classified in 3,650 synsets. Over 50% of these are derived from adjectives. 680 adverbial word forms are coded as antonymic pairs and approximately, 100 are coded for semantic domain. These relations constitute the sole inter-syntactic category connections for WordNet adverbs.

3.2.4 Inadequacies of WordNet

Although the design and implementation of the knowledge base is theoretically and empirically grounded, WordNet has a number of well-documented problems. A primary inadequacy is the issue of coverage. The coverage of WordNet is extensive but by no means complete. Indeed, complete coverage of all words and concepts in any language may be beyond the scope of any knowledge base, particularly one, such as WordNet, that is updated by hand. A related issue is the issue of coverage across all types of English. The current implementation of WordNet is based on American English. This focus

results in the omission of common terminology from other types of English and, in some cases, the dense population of the knowledge base with respect to certain American institutions, such as baseball.

A further inconsistency lies in the treatment of proper nouns in WordNet. There does not appear to be a fixed criteria for inclusion or exclusion of proper nouns in the knowledge base. Natural Language Applications that aim to treat proper nouns in some consistent fashion cannot rely on the presence *or* absence of these instances of proper nouns within WordNet. Another criticism that is often levelled at the knowledge base regards the level of sense granularity. The distinctions between senses can be so fine-grained that even metaphorical uses of terms are encoded as separate synsets. Like the uncertainty regarding proper nouns noted above, researchers cannot rely on the presence or absence of metaphorical uses of terms in the knowledge base and this can adversely effect applications intended to tackle figurative language.

In summary, inconsistencies of coding of synset contents and relations in the knowledge base has created a structure which, while extremely useful as a real world knowledge source, is unevenly populated and incomplete. Chapter 4 sets out an in-depth structural analysis of WordNet with the aim of addressing the question of variant data sparseness in the knowledge base.

3.3 Corpora

This section describes the provenance and contents of the three news text corpora used in the evaluation of the Storylines system reported in chapters 5 and 6. The three corpora — the RTÉ Corpus, the Australian Corpus and the Google news corpus — are each described in turn.

3.3.1 The RTÉ Corpus

RTÉ and RTÉ Interactive

Radio Telefís Éireann (RTÉ) is the Irish Public Service Broadcasting Organisation. It is a statutory corporation which provides a comprehensive service on radio (since 1926), on television (since 1961) and on the Internet. One of the main components of the service provided by RTÉ, as the Irish National Broadcaster, is its news service. All current stories are stored in a news database. This database is a comprehensive software system to facilitate the gathering, collating, searching and updating of news stories.

The database contains all news stories logged by RTÉ reporters, all news wire sources from agencies, such as Reuters and Agence France Presse, and all re-edited versions for the various radio and television news programs. Hence, the database contains all news sources and copy generated over a period of days.[11] The internal organisation of the system, however, is dictated by the source or purpose of a given text rather than according to the text content. The bins of the database hold all the stories for a given news program or from a given source. Links between stories, according to content, are maintained in a slightly arbitrary fashion by file naming conventions which are not strictly adhered to. For example, two stories about the Moriarty tribunal might be named *Moriarty1* and *Moriarty2* but would be equally likely to be named *Tribunal* and *Moriarty*. This naming inconsistency is more pronounced when the same topic is news-worthy over a longer period of time. A story about a wrecked oil tanker on Day 1 of the incident might be named *Tanker*, on Day 2 *Prestige Tanker* and on Day 3 *Oil spill*. Such inconsistencies arise in particular when stories are written or edited by different individuals on

[11]The actual time a story is held depends on how busy the system is at any time. Periodically older stories will be dumped from the database as new ones take their place.

different days. Therefore, initiates of the system use their own expertise to find and follow new stories. For novices, the task is not so easy as stories are not organised or searchable on a theme or topic basis.

The RTÉ Interactive editing process

The Internet and teletext services are provided and maintained by the RTÉ Interactive division of the corporation. The website and aertel (the teletext service) provide up-to-the-minute information on current news stories, entertainment, sport, business and weather. The team at RTÉ Interactive run to a 24-hour schedule constantly updating the contents of the site.

The manager of the Interactive team at RTÉ at the time of collection, Bree Treacy, designates their task as one of repurposing: taking texts from some source and producing a text that is appropriate for the medium of either Internet or teletext. Ms Treacy notes that the style of Internet and teletext texts should be "shorter" and "snappier" than their broadcast counterparts. The Interactive editors are concerned with the visual impact of the text and its accuracy in the context of constantly changing news. At a basic level, the editing process involves correcting typing errors and reformatting sentences and paragraphs to the style of Internet or aertel stories. The editors also write the headlines and includes links or images as required. As to the information content of the texts, editors may condense the information to fit the desired media or update information that has become out-of-date or was incorrect.

The Texts

The corpus of RTÉ texts consists of 628 individual texts gathered by the researcher over 5 days on site at RTÉ Interactive in Donnybrook, Dublin:

22 and 27 November and 4, 5 and 6 December 2002. The texts consist of the news articles posted on the RTÉ interactive website over the course of each day, their sources and their updates throughout the day. The sources are, for the most part, either RTÉ news copy or Reuters or AFP reports. A minority of stories are based directly on garda reports from the RTÉ news database. The texts are grouped by topic area and date although story topics for different days are not mutually exclusive. Appendix A.1 lists the story topics for the complete corpus of 628 texts. A subset of 25 texts was selected for use in the gold standard experiment. These texts are printed in full in appendix A.2.

3.3.2 Australian Corpus

The Australian Corpus of news texts was very kindly donated by Dr. Michael Lee at the University of Adelaide, Australia. The corpus consists of 50 documents from the Australian Broadcasting Corporation news mail service which provides e-mails of the headline news stories of the day. The texts are approximately 80 words each in length. This corpus was used in an experiment carried out by a group of researchers at the University of Adelaide to elicit a gold standard of human similarity judgments of news texts. Details of this experiment are set out in detail in Lee *et al.* (2004) and recapitulated in section 8.2.3 of this thesis. The fifty texts are set out in full in Appendix A.3.

3.3.3 Google News Corpus

The Google News website is an on-line news service which presents the key news topics of the day. For each news topic, there is a selection of links to on-line newspapers around the world. The site, therefore, gives users an overview of how a particular topic is treated in the world media on a

particular day. The task of finding and grouping similar news articles is carried out "entirely by a mathematical algorithm, based on many factors including how often and where a story appears elsewhere on the web".[12] The small corpus of news text used in the differential content identification experiment described in section 8.6 was collected from the Google News site. I implemented a perl script to trawl Google News for links, download stories grouped by topic and strip the documents of HTML mark-up. A subset of sixteen news articles to be used in the experiment was hand-picked from among these automatically down-loaded articles as the Google News algorithm is not 100% accurate occasionally resulting in unrelated articles being grouped together. The sixteen texts were grouped into eight text pairs on the same topic which are printed in full in Appendix A.4.

3.4 Summary

This chapter described the materials that have been used during the course of this research in the computational applications and evaluation processes. Section 3.2 outlined the key organisational principles of the WordNet knowledge base. Chapter 4 expands on this discussion providing an in-depth analysis of the structure of the knowledge base. Section 3.3 set out the origin, contents and compilation process for the three news text corpora that are used to evaluate the Storylines text representation and differentiation system reported in Chapters 5 and 6 of this thesis.

[12] From personal correspondence with the Google News team.

Chapter 4

The Topology of WordNet

4.1 Introduction

Chapter 3 outlined the organizing principles and design of WordNet. This chapter discusses aspects of the dimensions and shape of WordNet, viewed as a graph structure and what this may reveal about concept definition and concept specificity in the database.[1] Much work has been done on the notion of semantic relatedness between nodes in WordNet.[2] This chapter addresses a similar question — how comparable are two synsets in the WordNet network, not in terms of their content but in terms of the level of granularity or specificity they represent. This addresses one of the core aims of the thesis set out in section 1.1: to devise a means of factoring out the idiosyncrasies of a knowledge base, in this case WordNet, in order to fully exploit the potential of that knowledge base in Natural Language Processing applications, in this case the automatic estimation of text similarity and difference.

Although WordNet is a substantial knowledge base, as noted in section 3.2.3, it is not comprehensive. I do not know of work that records comparisons with arguably comparable resources like that supplied by CYC (Lenat and Guha 1990), however I expect that variant sparseness of coverage is endemic to all comparable knowledge bases. The level of detail in certain domains is essentially an accident of production dependent on several factors such as the day, the lexicographer and the level of interest.[3] Applications that use the data in WordNet to carry out some NLP task may themselves be subject to its vagaries. For example, two towns of comparable size in Ireland, Limerick and Drogheda, Limerick is encoded as both a port city and

[1] The content of this chapter has been published in (Devitt and Vogel 2003a).

[2] See (Budanitsky 1999) for a comprehensive survey of relatedness measures relating to WordNet and other knowledge bases, such as Roget's Thesaurus.

[3] For a case in point, note the recent addition of numerous concepts related to terrorism in WordNet 2, given the current political climate.

a type of poem whereas Drogheda is encoded as a battle, being the site of a 16th century battle. A topic identifier using WordNet as its knowledge base might identify texts about Drogheda to be historical or military, without the second possibility of the topic relating to modern day Ireland.

The aim of this chapter is to record statistics about WordNet, version 2, that may contribute to defining a notion of specificity that is determined by the topology of WordNet, and sensitive to variance in coverage across topic areas. The measures are applicable to any knowledge source that has a definable topology. The measures reported here are based on the noun and verb knowledge bases alone as the adjective clustering structure is inherently different and as such does not lend itself to some aspects of this kind of analysis.

This first section comprises nine subsections each detailing some basic measures for WordNet that characterize its overall topology: graph and node type § 4.2.1, taxonomic distribution § 4.2.2, parentage § 4.2.3, multiple hierarchy overlap § 4.2.4, node degree § 4.2.5, depth and height § 4.2.6, coordinate terms or sisters § 4.2.7 and clustering coefficients § 4.2.8.[4] Section 4.2.9 discusses polysemy and how the addition of polysemy links radically change the shape of the WordNet graph structure (Sigman and Cecchi 2002). This is followed by a description of an experiment to determine human judgments of node specificity in section 4.3. Finally, section 4.4 sets out some conclusions regarding how these structural features may be combined to estimate a metric of node specificity in WordNet. This detailed structural analysis of the knowledge base, the empirical investigation of the concept of node specificity in WordNet and the application of structural features to estimate

[4]These features were calculated from the WordNet graph using a perl script I implemented.

a specificity metric for nodes in the knowledge base constitute significant contributions of the thesis. The application of these measures to the task of automatic similarity estimation is described and evaluated in chapters 5, 6 and 8.

4.2 Topological features

4.2.1 Some basic measures

As this section deals with the *structure* of WordNet rather than its content, we refer to WordNet and its synsets in terms of a directed acyclic graph and not a tree as it allows multiple inheritance. Henceforth, we use the terms "node" and "synset" interchangeably.

Of 79,689 nominal synsets or nodes in WordNet 2, 58,586 or 78.65% are leaf nodes, leaving 15,902 internal nodes. Likewise, in the verb troponym hierarchy, approximately 76% of nodes are leaf-nodes. Furthermore, these leaf nodes may not be distributed evenly throughout the knowledge base. A χ^2 test of independence was carried out to determine whether the distribution of leaf to non-leaf nodes was the same across the nine sub-hierarchies of the noun knowledge base. The null hypothesis for this test is as follows:

> H_0: there are similar proportions of leaves in each of the sub-hierarchies (i.e. their distribution is not significant).

The results of the test, however, provide strong evidence for rejecting this hypothesis. The difference in the distributions of leaf versus non-leaf nodes across the nine noun sub-hierarchies was found to be statistically significant, $\chi^2(8, N=75180)=1126.98$, $p \leq 0.001$. Analysis of particular measures across WordNet, such as height and branching factor, must take account of the fact

that leaves are not distributed evenly throughout the knowledge base and that this factor and, more particularly, the predominance of leaf nodes in WordNet may skew results for other structural measures.

4.2.2 Dimensional distribution

Table 3.5 set out the nine designated most general root nodes to dimensions of the noun taxonomy. Bar chart 4.1 illustrates the distribution of nodes within these hierarchies. Nodes clearly follow an exponentially decreasing Zipfian distribution. The Entity hierarchy is by far the largest and as such merits some investigation as a separate unit. This is concrete evidence of one aspect of the variance mentioned in §4.2.1.

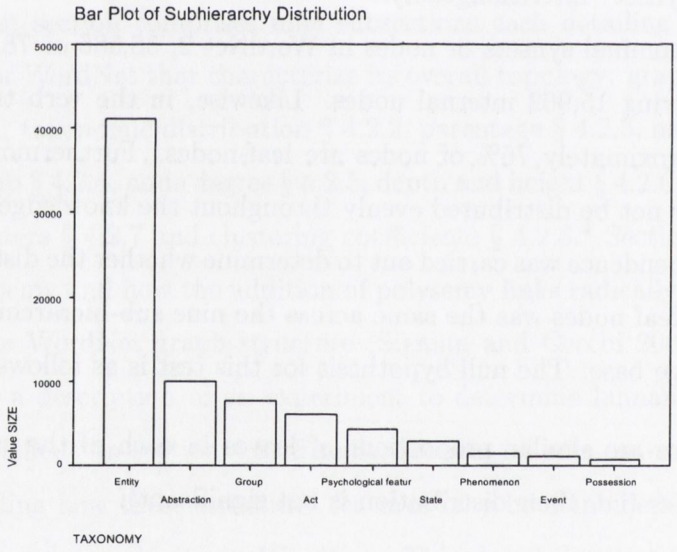


Figure 4.1: Bar chart of synset distribution in top noun hierarchies

As discussed in section 3.2.3, the verb and noun components of WordNet are structured according to quite different criteria. This is reflected in the

differences in the sub-hierarchies of the two components. There are over five hundred top nodes in the verb structure. Of these, 223 are not connected by a troponym relation to any other synset. The other synsets constitute the top of hierarchies of widely varying sizes, the largest stemming from the "change" synset, containing 2,032 synsets.

4.2.3 Multiple inheritance quantified

As noted above, the taxonomy does allow multiple inheritance. Although, the average number of parents is approximately one, the maximum reaches five. For example, the node referring to the multi-talented "Harley Granville-Barker" inherits from the more general nodes: "actor", "critic", "theatre producer", "director" and "playwright". Similarly, the more general "sphere" and "model" nodes are parents of the synset representing "globe". In all, these multiple inheritance nodes amount to just 2.28% of the total taxonomy.[5]

The histogram in figure 4.2 shows the distribution of nodes with more than one parent according to their depth in the hierarchy. Section 4.2.6 discusses the normal distribution of nodes in WordNet according to their depth in the graph, illustrated in figures 4.3 and 4.4. Figure 4.2 below strongly suggests that multiple inheritance nodes are also normally distributed following the distribution of nodes at different depths in WordNet and, thence, their

[5]In an inheritance hierarchy of verbs, constructed using the troponym relation, only 31 nodes have more than 1 parent. In fact, 2 parents is the maximum. In effect, multiple inheritance is a more prominent feature of nouns. This is to be expected as they lend themselves to a hierarchical structure more readily than verbs. If entailment and troponym relations are used to build the verb hierarchy, 2.5% of nodes have multiple "parents". However, whether this hybrid parentage is comparable to parentage in a single-relation inheritance hierarchy is debatable.

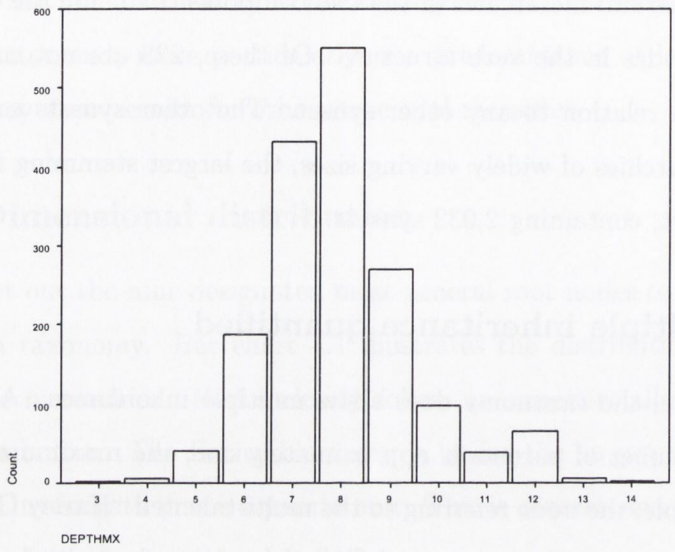


Figure 4.2: Histogram of depth for noun nodes with multiple parentage

effects propagate down the hierarchy.

One might expect, in line with Touretsky's (1986) definition of inheritance systems, that multiple parentage would imply a more specific concept node, from a content point of view. One might also posit that nodes deeper in the hierarchy are more specific. In this case, synsets in the right tail should be of comparable high specificity. Content inspection reveals the following as a sample of the highly-specific concepts in the right-tail of the distribution.

(4.1) sea bass (parents: serranid fish and food fish)
cytology (parents: biology and microscopic anatomy)
self-condemnation (parents: confession and accusation)

While nodes in the left tail, though with multiple parents, are less specific due to their position in the hierarchy

(4.2) person (parents: organism and causal agency)
artefact (parents: object and whole)

This contrast in generality of concepts, however, may be due solely to depth in the hierarchy rather than multiple parentage. Example 4.3 below are synsets deep in the hierarchy with single parents. They may be deemed equally specific to examples 4.1 above.

(4.3) beagle (parent: hound)
palomino (parent: horse)

On the other hand, examples 4.2 and 4.4 are at the same depth in the hierarchy but the latter, single parent nodes, could be deemed more specific.

(4.4) damnation (parent: denunciation)
office (parent: occupation)

This might suggest that depth is a stronger indicator of specificity than multiple parentage and that the interaction of the two factors is not clear cut. It is possible that multiple parentage near the top of the graph indicates greater generality while the inverse may be true closer to the bottom. One could also posit that it is depth and not multiple parentage at all that has an effect on perceived specificity.

A second analysis of instances of multiple parentage in WordNet was carried out to determine the distribution of such nodes across the nine noun sub-hierarchies. Table 4.1 sets out the contingency table for a χ^2 test of independence of multiple inheritance across the sub-hierarchies. The null hypothesis for this test is as follows:

H_0: the proportion of nodes with more than one parent is similar across the sub-hierarchies.

The results of the test provide strong evidence to reject this hypothesis, $\chi^2(8, N=75180)=324.27, p \leq 0.001$. This would suggest that, although multiple inheritance nodes are normally distributed throughout the depth of WordNet, their distribution is significantly different across the different subhierarchies. This would strongly support the hypothesis that multiple inheritance is significantly more prevalent in certain sub-hierarchies than others, as illustrated by contingency table 4.1.

	Entity	**Psych**	**Abs**	**State**	**Event**	**Act**	**Group**	**Poss**	**Phen**
Overall	39708	4147	9898	2770	1049	5992	7644	690	1417
> 1 parent	1365	77	130	102	12	83	69	8	28

Table 4.1: Contingency table for multiple parentage, χ^2**(8, N=75180) = 324.27, p<0.001**

4.2.4 Hierarchy Overlap

It should be noted that multiple inheritance does not entail an overlap across noun sub-hierarchies. In fact, only 2,072 noun synsets belong to two distinct subhierarchies. The number of synsets in the subhierarchy intersections is set out in table 4.2. Some examples of synsets belonging to two hierarchies are listed below:

- abstraction & event: musical
- group & event: parade
- entity & abstraction: writing
- abstraction & group: table of contents
- state & psychological feature: motion sickness

Of the 2,072 overlapping synsets, the following forty-three inherit from more than two subhierarchies:

- Entity, Psychological feature, Abstraction: 1 synset (hoodoo)
- Entity, Psychological feature, State: 7 synsets referring to saints
- Entity, Abstraction, Act: 3 synsets relating to invitations
- Entity, Event, Act: 1 synset (Actium)
- Event, Act, Group: 11 synsets such as parade and cavalcade
- Psychological feature, Abstraction, Act: 4 synsets denoting types of *a cappella* singing (e.g. "gospel singing")
- Psychological feature, Abstraction, Act and Event: 16 synsets, "stage dancing" and its hyponyms

While hierarchy overlap may be an interesting feature it accounts for just over 2% of the noun synsets in WordNet 2 and as such is not a generally applicable measure.

4.2.5 Branching Factor

The measure of node degree or branching factor here assumes the notion of dominance. It is calculated according to equation 4.5.

$$BranchingFactor = \#\ of\ Descendants + 1(the\ node\ itself) \qquad (4.5)$$

This avoids problems with zero values in subsequent metrics and corresponds to a common definition of dominance as a reflexive relation (Partee *et al.* 1993).

	Psych	Abs	State	Event	Act	Grp	Poss	Phen	Total
Ent	59	201	63	1	30	250	1	2	580
Psy	—	137	34	17	172	0	0	24	443
Abs	—	—	9	87	56	56	10	2	558
State	—	—	—	3	0	0	0	1	111
Ev	—	—	—	—	947	12	0	9	1076
Act	—	—	—	—	—	11	0	11	1127
Grp	—	—	—	—	—	—	0	0	329
Poss	—	—	—	—	—	—	—	0	11
Phen	—	—	—	—	—	—	—	—	50

Table 4.2: Hierarchy overlap count

Branching factor (BF) for nouns in WordNet ranges from 1 to 573 with an average value of 2.023. Excluding leaf nodes (i.e., BF=1), however, the average branching factor value rises to 5.793. Indeed, 75.8% of internal 15,902 nodes have a branching factor of less than five and almost 97% a value of less than twenty. Despite the different organising principle for verbs in WordNet, branching factor in the verb hierarchy is very similar to the noun hierarchy. It follows a Zipfian distribution. The range is from 1 to 393. Excluding leaf nodes, the average value is 4.97. Again, 79.7% has a value of five or lower and 97.8% of values are below twenty.

The distribution of BF values is the same across all the sub-hierarchies of WordNet: a Zipfian or power-law distribution decreasing exponentially. This suggests that overall, in all subhierarchies, the structure is not shallow: low branching with a large number of total nodes suggests greater overall depth in paths. This analysis of branching factor provides an insight into the structure of WordNet. If the measure proves to be indicative of knowledge

base structure, it could be employed, in combination with other measures, for example depth and parentage, to estimate population density of nodes in WordNet. In the following section, I explore the features of depth and height in the noun and verb systems with aim of determining whether these measures are indicative of interaction between node information content and graph structure in WordNet.

4.2.6 Depth and Height

As each node may be parent to or descendant of several lineages, nodes may have several possible values for both height and depth. The values discussed here are

- Maximum depth: longest path from node to a top taxonomy node
- Minimum depth: shortest path from node to a top taxonomy node
- Maximum height: longest path from node to any leaf node
- Minimum height: shortest path from node to any leaf node

The distribution of depth values among noun synsets in WordNet, whether maxima or minima, is normal (see figures 4.3 and 4.4). The data excluding leaf nodes is not substantially different. The means differ by 0.5 (7.1 with leaf nodes, 6.6 without) but the distribution is comparable.

The data for height, however, displays the effects of the preponderance of leaf nodes in the taxonomy. The maximum distance from any node to a leaf node is 5. Two-thirds of all internal nodes are a single node from the bottom of the taxonomy and 93.6% of nodes are a mere 1 or 2 nodes from a leaf node. In fact, for all values of the minimum height variable, the distribution of the depth variable is normal indicating that leaf nodes are spread evenly

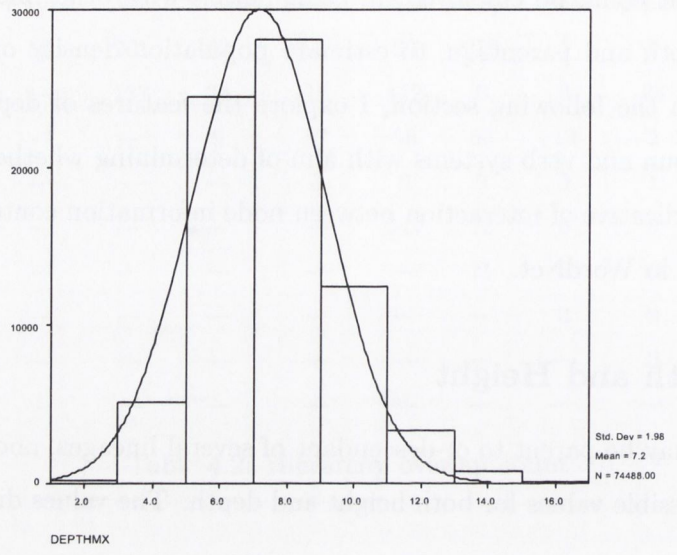


Figure 4.3: Nouns: Histogram of Maximum Depth

throughout all levels of the graph. Figures 4.5 and 4.6 show that, rather than correlating with depth values, both maximum and minimum height values follow the same distribution as words in natural language: a Zipfian distribution (Zipf 1935). A final analysis of depth and height in the nine noun sub-hierarchies, outlined in table 4.2.6, suggests that the distribution of depth and height maximum and minimum values is similar within each hierarchy as within WordNet as a whole.

This analysis of depth and height within WordNet suggests further hypotheses for node specificity estimation based on structural features. Given the distribution of measures of height in WordNet, it would seem that depth would be a better measure of specificity. A minimum height value of two does little to suggest how specific a concept may be, for within this selec-

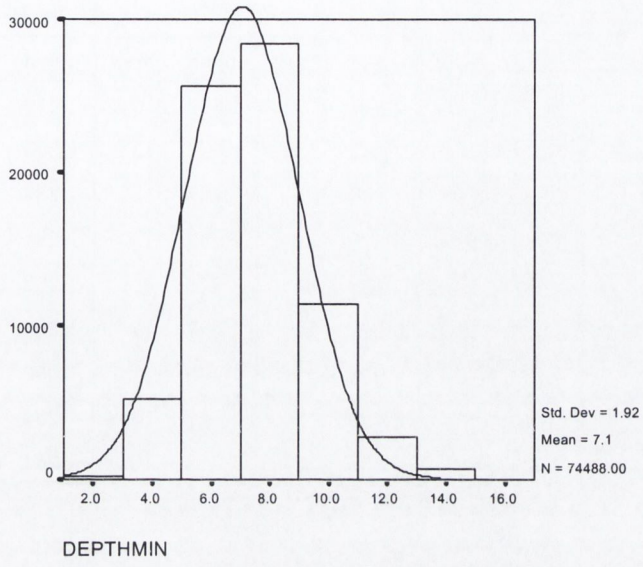


Figure 4.4: Nouns: Histogram of Minimum Depth

tion are the following sample nouns: *production, voodoo, group, refracting telescope, citizenry and floor.* The broader range and normal distribution of depth values would suggest that this measure is more indicative of WordNet structure and, therefore, could prove to be a good predictor of node specificity in nouns. In the verb system, the distributions of depth and height values are similar to the nominal distributions, that is, normal and Zipfian respectively. Verb synsets, however, form a significantly shallower structure with the average depth under half that of the noun hierarchy at 3.53. This would suggest that, while depth may be a good indication of specificity among nouns, among verbs it may not be as informative an indicator.

	Max Depth		Min Depth		Max height		Min Height	
Hierarchy	mean	max	mean	max	mean	max	mean	max
Entity	7.651	18	7.514	15	1.339	16	1.219	5
Psychological feature	7.294	14	7.184	13	1.443	13	1.263	3
Abstraction	7.392	15	7.264	15	1.447	15	1.28	4
State	5.968	12	5.879	12	1.447	15	1.28	4
Event	5.287	10	5.152	9	1.448	9	1.282	4
Act	6.095	12	6.01	12	1.497	12	1.294	5
Group	6.093	12	6.053	11	1.149	11	1.092	4
Possession	6.449	13	6.429	13	1.577	13	1.284	3
Phenomenon	5.925	10	5.886	10	1.428	10	1.281	3

Table 4.3: Depth and height values in sub-hierarchies

4.2.7 Coordinate terms: node sisters

As the number of leaf nodes by far exceeds the number of internal nodes, it may be more pertinent to discuss connectivity in terms of node sisters than node children. The statistics for nouns and verbs, in this case, is similar. In the noun hierarchy, the sister variable ranges between 0 and 798. The average value is 44.85 with a standard deviation of 93.68. 88% of all nodes have a value below 100. Likewise, for verbs, the variable ranges between 0 and 578 with an average of 52.88. The variable is not normally distributed in either knowledge base but rather follows a Zipfian distribution. The highest values for the sister variables occur in the middle of the taxonomy, from depth 6 to 9. This might indicate that some core concepts occur just above this level, for example, "city" or "author", which may form the centre of semantic clusters. The broad range of values and Zipfian distribution and possible tendency for nodes with similar numbers of coordinate terms to represent clusters in the

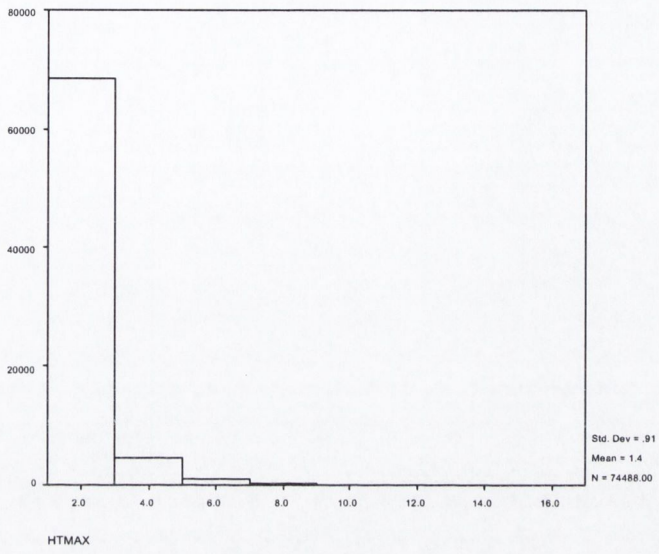


Figure 4.5: Nouns: Maximum Height (distance to any leaf node)

knowledge base suggest that this feature may indicate different aspects of WordNet structure than other features such as depth or branching factor. For this reason, it could be employed in specificity estimation to complement or counter-balance these features.

4.2.8 Clustering coefficients

Clustering coefficients as a fine-grained measure of graph topology and connectivity have been posited by Watts and Strogatz (1999). This coefficient measures the relative number of connections between neighbouring nodes in a network. It is, therefore, a measure of how clustered an area of a network may be. The formula to calculate the clustering coefficient, C_i, of a node i

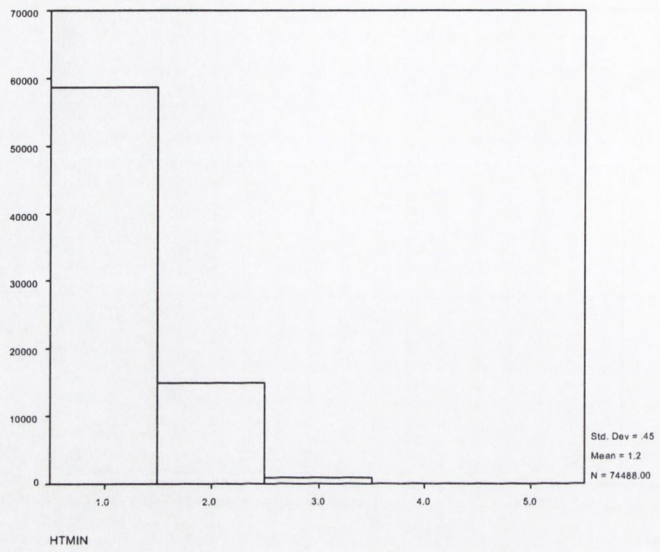


Figure 4.6: Nouns: Minimum Height (distance to any leaf node)

is set out in equation 4.6.

$$\frac{2\Sigma_i}{k_i(k_i - 1)} \tag{4.6}$$

In this equation k_i is the number of connections from a node i to its neighbouring nodes and Σ_i is the number of connections between those k_i nodes. Higher-order coefficients measure connectivity between a node's immediate and more distant neighbours to a given distance. The coefficient gives a normalized measure of connectivity across a whole graph.

A first point to note is that the basic cluster coefficient is not useful for a graph such as WordNet. Only sixty-two synsets have a coefficient higher than zero. This would indicate that the nodes in WordNet do not form strong clusters readily. This is most probably due to the hierarchical rather than network structure of the taxonomy. The higher order measure, tak-

ing immediate neighbours and nodes at one extra remove, is a more useful value, particularly for internal nodes where the distribution is normal and the mean is 0.337. Somewhat surprisingly, this is also true of verbs where the hierarchical nature of the taxonomy is not the dominant feature. The cluster coefficient was calculated including troponymy and entailment relations. Even still, 99.5% of verb nodes had a value of 0 for the first order measure. Again, the second order coefficient was more useful, ranging from 0 to 0.17 but still the vast majority of nodes have a clustering coefficient of 0.

This analysis would suggest that WordNet is not tightly clustered but that its nodes may form clusters of wider diameter. It is not clear that these coefficients would be useful in estimating a measure of node specificity for two reasons. Firstly, for most nodes, the coefficients have no value and, as such, they are only marginally representative of WordNet structure. Secondly, the calculation of the coefficients is based on the branching factor, sisters and parentage values and, therefore, it is not clear what *extra* information the coefficients would contribute to a specificity estimation algorithm.

4.2.9 Polysemy in WordNet

Many word forms in WordNet belong to more than one synset. This section looks at how prevalent polysemy is in WordNet and discusses the impact of polysemy relations on the WordNet graph. WordNet contains 147,390 noun, verb and adjective word forms. Each word form belongs to at least one and at most fifty-nine synsets. Table 4.4 sets out the range and mean of values for polysemy among synsets of each syntactic category in WordNet. What is particularly interesting about this variable is the effect it has when added to the basic inheritance hierarchy defined by semantic relations in WordNet. Sigman and Cecchi (2002) state that the inclusion of polysemy

Syntactic category	Range	Mean	Word forms	Synsets
Noun	1-32 (head)	1.236	114648	79689
Verb	1-59 (break)	2.179	11306	13508
Adjective	1-29 (heavy)	1.447	21436	18563

Table 4.4: Polysemy values for nouns, verbs and adjectives

relations between nodes of the graph convert the structure from a graph with low connectivity to a "small world" network. Small world networks are characteristic of many networks found in nature and technology from the world wide web to the spread of infectious diseases (Watts and Strogatz 1999). Watts and Strogatz hold that:

> Models of dynamic systems with small world coupling display enhanced signal propagation speed, computational power, and synchronizability (Watts and Strogatz 1999, p.440).

It is heartening, therefore, that the mental lexicon, which is a high-speed, powerful processing system, should be modelled by a knowledge base which displays this characteristic high connectivity. As noted in section 4.2.8, the graph constructed of semantic relations between synsets alone does not display these properties. Perhaps, a structure which includes lexical polysemy relations, thereby creating hubs of central concepts in the graph, may be more appropriate to a representation of the lexicon and cognitive processes connected with it. The graph representation which I designed that is the basis of the similarity estimation system reported in chapter 5 models this structure. To date, the added polysemy dimension of the structure of WordNet has not been exploited in the task of text similarity estimation. However, one of the

projected extensions of this research, set out in section 9.2, is to develop this aspect of the system in order to address the issue of lexical ambiguity in text.

4.3 Specificity experiment

The measures discussed in the previous section aim to characterise the topology of WordNet in order to establish a basis for comparison of nodes based not on content, which cannot yet be quantified computationally, but on their position within the knowledge base, which can be determined automatically. Nodes of similar generality or specificity, and hence perhaps information content, could thereby be identified. A notion of generality or specificity is central to this approach. For this reason, a pilot study was undertaken to determine whether degree of specificity or generality could be deemed a characteristic of concepts, as represented by Synsets in WordNet. It is with the aim of defining an independent measure of concept specificity to factor out effects of the over— or under—population of the knowledge base that the structural analysis of WordNet reported above and this empirical study were undertaken.

4.3.1 Experiment Goals

The study was intended to answer three questions:

1. Do people have a notion of generality/specificity for concepts?
2. If so, do people agree on what concepts are more/less specific?
3. What features of WordNet do these judgments correlate with?

The first two questions aim to identify whether specificity is a useful concept and one on which there is consensus. The third question is intended

to justify the use of topological features of WordNet as part of a *semantic* representation of texts.

4.3.2 Methodology

The study was designed as a small-scale pilot. Initial choices of materials and presentation format were refined after two talk-aloud protocols conducted with a single volunteer.

Participants

A small number of participants were solicited from among personal contacts of the researcher. There were 11 participants in total in the age-group 30-40. The group consisted of six females and five males. All are native speakers of English. One of the group is left-handed.

Materials

Fifteen sets of three noun synsets drawn from WordNet were selected.[6] The synsets were presented as the synonymous word set and the explanatory gloss from WordNet. For example:

(4.7) trout – (any of various game and food fishes of cool fresh waters mostly smaller than typical salmons)

As the experiment was intended to identify associations between specificity and WordNet topological features, the sets of synsets controlled for different features. The topological variables under investigation in the data were:

- depth

[6] All experiment materials are presented in Appendix B.

- branching factor
- number of sisters
- parents
- sub-hierarchy

The first ten sets of synsets are all drawn from the entity hierarchy. In the first three sets, the values of all variables were held constant as a control. Branching factor is varied in sets IV, VII and VIII, depth in sets V, IX, X and sisters in set VI. In sets XI-XV, the synsets are drawn from different hierarchies. Where one variable is varied all other variables are held constant. Within the range of possible synsets that satisfied each experimental condition, the synsets were drawn at random.

Procedure

The experiment was conducted as a questionnaire which was distributed to the volunteers and subsequently collected by the researcher. The questionnaire first defined how the "concepts" were presented in the experiment — as a set of words and a definition. The fifteen sets of synsets were presented as fifteen individual ranking tasks where respondents were asked to rank the synsets from most to least general. Respondents were informed that concepts they deem of equal generality may be ranked on the same level. If respondents did not think it was possible to rank the concepts, they were asked to mark the set with an asterisk. For each respondent, the specificity rankings were converted from a single tripartite ranking to three binary rankings, as in the example below. The following is a hypothetical tripartite ranking, where one is most general and three is least general:

1. entity
2. mammal
3. sea bass

This ranking would be converted to the following three binary rankings, where > represents "is more general than".

- *entity* > *mammal*
- *mammal* > *seabass*
- *entity* > *seabass*

In this way, the fifteen tripartite—ranked sets for each respondent are converted into forty-five binary rankings.

4.3.3 Results

In response to the first two experimental questions regarding people's idea of specificity, the answer would seem to be positive. Yes, people have a notion of generality applied to concepts and, yes, people agree on what is more or less general. In practice, the first question is answered implicitly by the second: If people agree on which concepts they deem more or less specific, then they must share some common notion of what it means for concepts to be specific or general. In order to determine the level of inter-rater agreement in this experiment relative to that expected according to chance, the Cochran Q statistic (Cochran 1950) was calculated on the binary rankings.[7] The null hypothesis for this test was as follows:

H_0: any agreement between respondents was due to chance.

[7]See chapter 7 for a full discussion of inter-rater agreement measures.

The Cochran Q statistic on the experiment data at 44 degrees of freedom is 165.859, which would suggest that the null hypothesis may be rejected at $p \leq 0.0001$.

The third experiment aim was to determine whether there is any correlation between WordNet topological features and a measure of concept specificity. The data suggest that depth is the strongest correlate for specificity. Overall, concepts which are deeper in the WordNet hierarchy were deemed more specific than concepts higher up in the WordNet hierarchy. Contrary to expectations, in this pilot study, branching factor did not appear to have any impact on people's generality rankings. The data does suggest, however, that the number of coordinate terms of a concept is negatively correlated with the specificity of concepts. In this experiment, concepts with more sisters were deemed less specific. Membership of a particular sub-hierarchy did not appear to effect generality rankings. Although, it did seem to increase the difficulty of the task. There were only five instances of the asterisk being used to mark unrankable sets. Three of these were in sets X-XV where sub-hierarchy is varied.

I must stress that it is only possible to draw tentative conclusions on the basis of these findings. In order to derive statistically significant results, a study on a much larger-scale should be conducted, comparing many more concepts in WordNet while controlling for different features of its topology. Such a study is beyond the scope of this thesis. The pilot, however, has proved useful in suggesting hypotheses, even though definite correlations cannot be established. It also suggests directions for future research, as outlined in section 9.2. As regards the research reported here, the study results provide some evidence to support the hypotheses regarding indicative features of the structure of WordNet made in previous sections. These hypotheses

constitute the basis for the node specificity weighting algorithm described in section 6.6.1.

4.4 Some Conclusions on WordNet and Node Specificity

This chapter provides a detailed analysis of aspects of the structure of the WordNet knowledge base. The metrics described each provide a particular perspective of the WordNet topology. Previous sections have outlined the contrasting distributions of depth and height, the related concepts of branching factor, sister nodes and cluster coefficients, the notion of multiple inheritance and its significance within the taxonomy. The analysis provided in this chapter highlights which metrics contribute most to defining key features of the structure of the knowledge base and which are less informative. The empirical pilot study, outlined in section 4.3, provides evidence to support the hypothesis that some of these topological features correlate with human intuitions of concept specificity.

The indicative metrics that have emerged during the course of the research presented here can be employed to help smooth the unevenness and sparseness of coverage in WordNet and endemic to all knowledge bases. Furthermore, a model of the topology of WordNet would be useful in guiding interpretation of its content, particularly for non-humans. The more information we have about the structure of the knowledge base in abstract, the more confident we may be in our exploitation of its resources. The structural analysis and empirical investigation presented here constitute a major contribution to this end. The resulting hypotheses for node specificity estimation

have been implemented as specificity formulae[8] in the system, reported in chapter 5, which I developed to build meaningful text representations from free text using WordNet as a real world knowledge resource. It is aspects of the topology of WordNet outlined in this chapter which the system exploits in order to estimate text similarity.

[8]See section 6.6.1 for an exposition and discussion of the node specificity formulae.

Chapter 5

Meaningful Text Representation: an Implementation

5.1 Introduction

There is a vast profusion of information in text form languishing on web servers and in databases across the world at this moment which could be put to use, transformed into knowledge, if only it could be found and assimilated. Google, for one, has made great strides in the location of relevant information. But beyond this problem space of document location, there still remains the question of how to access the content of a text automatically once it is found. The previous chapter discussed approaches to the problem of how the content of a text may be represented and applications to derive such representations automatically. These approaches range from unstructured "bag of word" models, see section 2.3, to more structured representations, see section 2.4 and 2.5, which can be realised automatically with more or less requirements and to varying degrees of success. This chapter describes Storylines, an implementation to construct from free text a meaningful representation whose internal structure is derived from the text and the WordNet knowledge base. Section 5.2 sets out the aims of this implementation. The four modules of the algorithm are described in sections 5.4, 5.5, 5.6 and 5.7. Section 5.8 details the data structures used to represent the model. Section 5.9 discusses how key attributes of the model may be exploited to estimate its information content. Throughout the chapter, the news story in example 5.1, from the RTÉ news interactive corpus on 27 November 2002, is used to illustrate different aspects of the system.

(5.1) 1. Strike by NI firefighters under way

2. Firefighters' strike is under way

The British Army has taken over firefighting duties in the North at the start of an eight-day strike by fire crews.

Almost 2,000 full-time firefighters have left their posts and pickets are being put

on fire stations.

The pay strike by firefighters throughout Britain and Northern Ireland began at 9am this morning after overnight talks failed to avert it.

The Fire Brigades' Union has accused the British government of vetoing an agreement that could have prevented the action.

The union said it had worked out a deal with employers that would have given firefighters a 16% pay increase, but the government had refused to help fund it.

5.2 Aims

Storylines, the implementation described in this chapter, strives for simplicity of application to yield complex, informative structure from unstructured text. The aim is to investigate the potential of this model while minimising pre-processing. The starting point is free text, without mark-up and without disambiguation.

The research reported in this thesis is carried out within the theoretical framework of lexical cohesion. Section 2.6.3 noted that many implementations based on the notion of lexical cohesion, in particular lexical chaining algorithms, build representations of texts based solely on nouns (Hirst and St-Onge 1998, Barzilay and Elhadad 1997, Stokes 2004). Storylines aims to extend the conventional focus of these applications to other syntactic categories besides nouns, namely verbs and adjectives. As noted in section 3.2.2, the degree of connectivity between the synsets of different syntactic categories in WordNet has been markedly low in previous versions of the knowledge base. In version 2.0, however, there is some integration, and Storylines exploits this in order to reduce information loss in the transition from text to text-based structure.

StoryLines also aims to circumvent some of the flaws inherent in a hand-

coded knowledge base, such as WordNet, to which any implementation using such a database is subject. In particular, Storylines addresses the issue of the uneven population of the knowledge base. As discussed in section 4.1, the differential information content of WordNet concepts is difficult to evaluate given the over-population of some topic areas in contrast with the under-population of others. The topological analysis of WordNet and WordNet concept specificity, set out in chapter 4, provides a framework whereby the local idiosyncratic characteristics of WordNet can be mitigated by global features of the knowledge base.

5.3 From Text to Graph: Basic Algorithm

The basic algorithm proceeds in four steps from free text to structured representation. The four modules are presented in the flow chart in figure 5.1. All components of the Storylines system were designed and developed for the purposes of this research, with the exception of the text tagger and chunker used in Module 1 and the WordNet synset access function used in Module 3. The input text is first tagged for part-of-speech and chunked using the LT CHUNK shallow parser from the University of Edinburgh (section 5.4). All nominal, verbal and adjectival terms are then extracted from the tagged text (section 5.5). For each of these terms, all synsets are extracted from WordNet using the WordNet synset access function. For each synset, all associated synsets and relations are extracted from WordNet using a component I designed and implemented to traverse the WN graph extracting relevant information (see section 5.6). This information is formatted as a list of binary relations between synsets in WordNet. The final module of the algorithm takes this list of relations, a file of WordNet topological in-

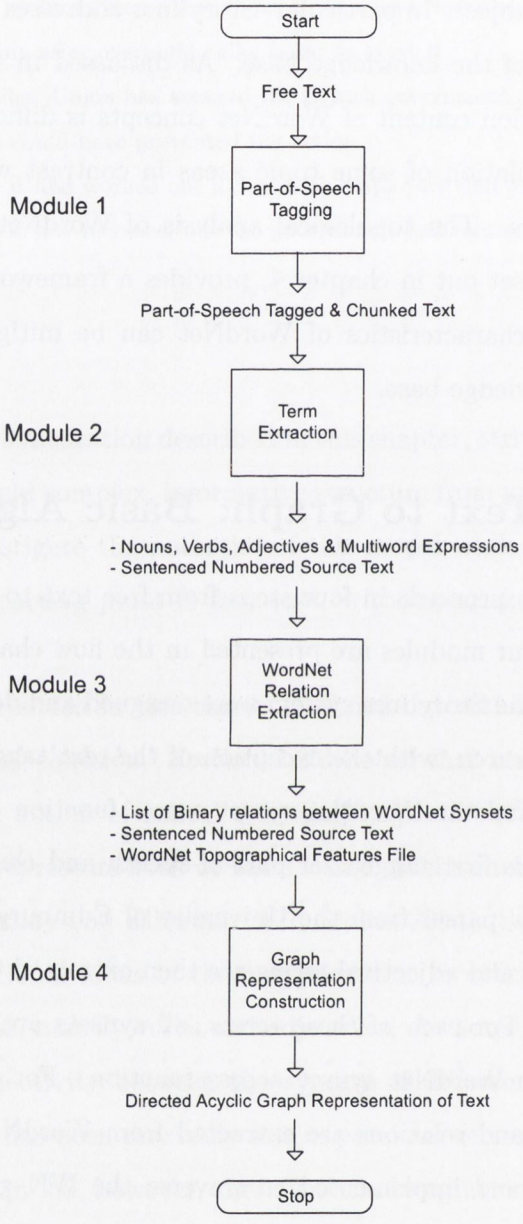


Figure 5.1: Storylines System Process Flow Chart

formation, generated off-line by a program I implemented, and the original source text as input and constructs a graph representation of the input text (see section 5.7). The following sections set out each stage of the process in turn. For each module, any theoretical issues that arise and resulting design decisions that were taken are discussed.

5.4 Module 1: Part of Speech Tagging

The input text is tagged using LT CHUNK, a tagger and chunker from the Language and Technology Group at the University of Edinburgh (Mikheev 1996, Finch and Mikheev 1997). The chunker takes as input free text and outputs text marked up for part-of-speech and syntactic chunk, as in example 5.2.[1]

(5.2) <S>Almost_RB [[2,000_CD full-time_JJ firefighters_NNS]] ((have_VBP left_VBN)) [[their_PRP$ posts_NNS]] and_CC [[pickets_NNS]] ((are_VBP being_VBG put_VBN)) on_IN [[fire_NN stations_NNS]]..._</S>

The chunker identifies simple noun and verb "groups" or chunks, with noun chunks delimited by square brackets and verb chunks by round brackets, as in example 5.2. The chunker groups sequences of determiners, pre-modifiers and nouns into nominal chunks. It does not provide an analysis of post-modifiers. Therefore, post-modifiers are not included in the same chunk as their head noun, as in examples 5.3 and 5.4.

(5.3) an agreement that could have prevented the action

[1]The part-of-speech tags used are a subset of the Penn tree bank tagset. See Appendix C.1 for the full list of tags and Appendix C.5.2 for the full sample tagged output for the Firefighters example.

[an_DT agreement_NN] [[that_WDT]] ((could_MD have_VB prevented_VBN)) [[the_DT action_NN]]

(5.4) the start of an eight-day strike

[the_DT start_NN] of_IN [[an_DT eight-day_JJ strike_NN]]

The verb chunks output by LT CHUNK are not verb phrases. They are composed solely of sequences of auxiliaries, modals, main verbs and some prepositions that have been analysed as verbal particles. The Language and Technology group report 96% to 98% accuracy for the tagger when all words are found in the tagger lexicon. The performance of the chunker is not quantified.

The first noun-based version of Storylines (Devitt and Vogel 2003b) did not include this initial tagging module. However, it was included in later versions that deal with verbs and adjective in text serving the function of a resolver of inter-category word sense ambiguity. This is not a requirement of the Storylines system but was intended to limit to some extent the dimensions of the final text representation and also to facilitate multi-word expression extraction. Although the tagger itself may introduce errors, up to 1 in 25 incorrect tags, it was deemed an acceptable risk in order to reduce the possible dimensions of fully ambiguous text.

5.5 Module 2: Term extraction

Algorithm 5.1 Module 2: Term Extraction

Input: *text* Part-of-speech tagged and chunked text (see appendix C.5.2)

Output: as follows:

1. *wordList*: list of nouns, verbs and adjectives and multiword expressions from the source text (see appendix C.5.3)

2. *sentences*: input text with each sentence numbered

```
1: for all sentence in text delimited by LT CHUNK do
2:   output sentenceNumber sentence minus tags to sentences
3: end for
4: for all chunk in text identified by LT CHUNK do
5:   output chunk minus tagging mark-up to wordList
6:   if chunk == noun then
7:     for all prenom such that prenom isa pre-nominal determiner or modifier do
8:       strip prenom
9:       output remaining chunk to wordList
10:    end for
11:    for all postnom such that postnom isa post-nominal modifier do
12:      strip postnom
13:      output remaining chunk to wordList
14:    end for
15:    for nounA in remaining n nouns in chunk do
16:      for nounB in n − 1 remaining nouns do
17:        output nounA .. nounB to wordList
18:      end for
19:    end for
20:  end if
21: end for
```

The second module involves extracting the nominal, adjectival and verbal

expressions from tagged text to be used as input to the rest of the system. This gives rise to the question of what should constitute an expression, term or word in Storylines. Halliday, in his discussion of forms of lexical cohesion, notes that:

> The concept of the lexical item .. is not totally clear–cut. (Halliday and Hasan 1976, p.292)

Given the difficulty of defining a lexical item, the issue arises of whether Storylines should deal only in "word" units, in the sense of orthographical units delimited by white space or whether multi-word expressions should also be included. Singleton (2000), on the mental lexicon, discusses the difficulty of specifying a firm concept of what is a word in the languages of the world. Furthermore, on the subject of multi-word expressions, he concludes that:

> even a conservative approach to the question of what counts as a lexical unit based on a criterion of semantic unitariness has to concede that there are lexical units which consist of more than one word. (Singleton 2000, p.57)

Storylines is intended to exploit the full potential offered by the WordNet knowledge base and as WordNet claims to provide a model of the mental lexicon, the Storylines term extraction algorithm takes Singleton's (2000) broad sense of lexical item which include words and multi-word expressions, such as 'Northern Ireland', 'public opinion', 'tax credit'.

If we accept that there are multi-word expressions in language, the question remains of how to identify these expressions in text. It was beyond the scope of this thesis to implement a comprehensive analyser for multi-word expressions. Therefore, Storylines extracts these expressions on the basis of the chunk boundaries marked by the LT CHUNK shallow parser. The types

of expressions that are handled are dictated by chunker output. For nominal expressions, these include any strings of determiners, pre-modifiers and nouns such as those set out in examples 5.5 to 5.7.

(5.5) vehicle registration tax

(5.6) public finances

(5.7) colourless green ideas

As the chunker does not include post-modifiers in nominal chunks, multiword expressions such as "Secretary of State" or "word of honour" will not be extracted. Storylines, however, addresses this question in a somewhat *ad hoc* manner by combining only those noun phrases separated by a single "of" preposition and applying the term extraction algorithm set out in 5.1 to the whole unit. This deals with the noun phrases in example 5.8.

(5.8) Secretary of State, Puss in Boots, supply and demand

The nominal multi-word expressions which are identified by the Storylines term extraction component are quite limited. It provides no analysis of idiomatic expressions such as 'over the moon', or 'down in the mouth'. Nor indeed is there an analysis of noun phrases with post-modifiers introduced by anything other than the preposition 'of', such as 'the man in the moon' or 'the house that Jack built'.

The verbal chunks identified by LT CHUNK are simpler in structure. Storylines, therefore, only extracts verbal multiwords that consist of a main verb with a verb particle, such as "blow up", "turn down". Storylines does not attempt to extract adjectival multiword expressions. As I note in section 9.2, inclusion of a more comprehensive multi-word analysis module could prove of great benefit to the system.

In summary, Storylines extracts all usable chunks provided by the LT CHUNK chunker. However, as noted in section 3.2.4, the coverage of WordNet which is the knowledge base for Storylines, is extensive but by no means complete. Therefore, although the Storylines term extraction algorithm identifies some multiword expressions, there is no guarantee that the knowledge base will contain these expressions, even if they are legitimate. For this reason, Storylines also extracts the components of multiword expressions for processing in subsequent modules. Example 5.9 sets out an example of a string of nouns that yields a list of noun permutations. Example 5.10 shows an identified multiword expression that contains two adjectives. The two adjectives are each returned separately but no adjectival multiword ("stable public") is posited.

(5.9) `vehicle registration tax` - `vehicle`
- `vehicle registration`
- `vehicle registration tax`
- `registration`
- `registration tax`
- `tax`

(5.10) `stable public finance` - `stable public finances`
- `public finances`
- `finances`
- `public`
- `stable`

The algorithm to extract single and multi word expressions from the tagged output of the LT CHUNK shallow parser is set out in algorithm 5.1.[2] The

[2]The algorithm is implemented as a Perl script. See Appendix C.2 for the source code.

output of this term extraction process, then, is a list of records, where each record has two colon-delimited fields, as in examples 5.11 to 5.13. The first field is a sentence number in the source text[3] and the second is a term located in this sentence.

(5.11) **Nouns**: 1:Strike, 1:firefighters, 2:firefighting duties, 5:Fire Brigades, 5:Fire Brigades' Union

(5.12) **Verbs**: 2:taken, 3:left, 4:failed, 5:accused, 5:prevented

(5.13) **Adjectives**: 2:eight-day, 2:firefighting, 3:full-time, 4:overnight, 5:British

The output terms consist of nouns, verbs and adjectives as well as multi-word nominal and verbal expressions.[4] These constitute the building blocks of the meaningful text representation built in Module 4 (see section 5.7). Within the framework of lexical cohesion, these terms through their inter-relations and interactions provide a means of accessing the semantics of a source text. The next section sets out how these inter-relations and interactions are identified using the WordNet knowledge base.

5.6 Module 3: WordNet relation extraction

Section 3.2 discussed the organisation of lexical items in WordNet into sets of synonymous words called synsets connected by a variety of lexical and semantic relations. Module 3 of the system I designed and implemented, Storylines, entails accessing the WordNet database to identify the synsets to which the

[3]The sentence number is used by the differential content component of Storylines described in section 6.8.

[4]See Appendix C.5.3 for the full wordlists extracted for the sample firefighters story.

terms extracted in Module 2 may belong. For each synset, the WordNet relations required in Module 4 of the Storylines program are extracted from the database. This component takes single word or multiword expressions as input and outputs a list of binary relations between synsets directly and indirectly related to these input terms, formatted as in example 5.14.

(5.14) @ 00098966n:00101927n 1 0 1 null Strike:0
#m 09444162n:07630155n 2 1 0 Firefighters:3 null

The sections that follow describe how this output is generated and what is its significance. Section 5.6.1 gives an overview of the Module 3 algorithm. Section 5.6.2 provides an in-depth analysis of how the semantics of the different WordNet relations dictate the extraction of synsets and relations from the knowledge base.

5.6.1 Algorithm overview

Synset extraction

The basic procedure for this module is set out in algorithm 5.2.[5] Firstly, the complete WordNet database files are loaded as a complex hashtable where the keys are the synset offsets and the values are an array of relation types,[6] such as hypernymy and cause, pointing to other synset offsets. Processing the input word list begins by tokenising the input into sentence number and term. For any term that is not one of a list of pronouns, the list of all WordNet synsets to which the term belongs is compiled. This step employs a system

[5]The algorithm is implemented as a Perl script. See Appendix C.3 for the source code.

[6]See section 3.2.2 for an in-depth discussion of WordNet relations and section 3.2.1 for a discussion of synset contents and an explanation of a synset offset.

Algorithm 5.2 Module 3: WordNet Relation Extraction

Input: *wordList*: List of words and multiword phrases (see appendix C.5.3)

Output: *relList*: List of binary relations between WordNet synsets (see appendix C.5.4)

Load WordNet as hashmap of synsets and relations between synsets
Read in *wordList*
for all line L *in wordList* **do**
  tokenise L into term T and sentence number *sent*
  if $T \in PronounList$ **then**
    do simple "pronominal resolution"
    output *relList* of relations between pronoun synset and related synsets
  else
    get *synsetList* for term T
    for all synset $S \in synsetList$ **do**
      get index i_T of term T in synonymous word list of synset S
      findRelations($S : i_T$, *sent*, $inTextFlag$)
      output *relList* for synset S and related synsets
    end for
  end if
end for

call from the Storylines program to the WordNet command line interface.[7] For each synset, the index of the source term in its list of synonymous words for that synset is also added to a list. This is to ensure that the system keeps track of the exact source words associated with each synset for polysemy handling in Module 4. For terms in the list of pronouns, the system call to WordNet is by–passed and an appropriate synset is assigned for each pronoun. For example, "he" is replaced by the synset for "male person", "you" is replaced by the synset for person. Subsequent sections make the distinction between synsets that are "present in the text" and those "derived from the text". The synsets which have been extracted at this first step by pronominal resolution or using the WordNet system call are flagged as synsets *present in the text.* This is not making any claim that synsets and terms are interchangeable, it is merely for ease of reference to make a clear association between the terms that *are* in the text and the synsets that represent those terms.

Relation Extraction

The findRelations process recursively generates the list of binary relations between synsets present in the text and synsets derived from them. The output consists of the following seven tab-delimited fields:

1. WordNet relation symbol;[8]

2. Synset1:Synset2: two synset byte offsets[9] divided by a colon;

[7]The interface is used directly for expedient morphological analysis as the WordNet library of access functions includes a comprehensive morphology component (Fellbaum 1998, p.124).

[8]See section 3.2.2 for a list of WordNet pointer relations and their symbols.

[9]See section 3.2.1 for an discussion of synsets and an explanation of the synset offset.

3. the source sentence number from which the synset binary relation derives;

4. inText1: a flag indicating if Synset1 is present in the source text;

5. inText2: a flag indicating if Synset2 is present in the source text;

6. term1: the source text term and synset word list index for Synset1, or null if the synset is derived from the text;

7. term2: the source text term and synset word list index for Synset2, or null if the synset is derived from the text.

As well as the WordNet relations, for each list of synsets referring to a single polysemous term, Storylines outputs a "pol" relation identical in format to the WordNet relations.[10] Example 5.15 sets out sample output for the firefighters story example.[11]

(5.15) @ 13788338n:13808178n 1 0 1 null NI:0
#s 13808178n:13836665n 1 1 0 NI:0 null
+ 13808178n:01355211v 1 1 0 NI:0 null
pol 01173759n:00919554n 1 1 1 Strike:0 Strike:0
pol 04679683n:13000678n 1 1 1 way :3 way :1

This final recursive process gives rise to three questions:

- which binary relations does the system output?

[10]Mani and Bloedorn (1999) developed a summarisation tool based on a graph representation combining WordNet synonymy and text-based phrase relations. The inclusion of an extra relation between all synsets in the same phrase or sentence was implemented. However, it produced an explosion of relations that led to an unjustified increased cost of computation and was, therefore, removed. Instead, inter-phrase connection is logged in the output in the sentence number field.

[11]The full output from Module 3 for the firefighters example is set out in Appendix C.5.4.

- how does the system define "association" with synsets present in the text?
- how and why is the process recursive?

Section 5.6.2 discusses these questions in relation to the literature about relation type, directionality of relations and relation distribution in WordNet.

5.6.2 Required WordNet relations

The representation built in response to input texts by Storylines, which is described in detail in section 5.7, is, in essence, a subset of the WordNet directed acyclic graph. The synsets present in the text constitute leaf nodes of this sub-graph. The remaining nodes in the sub-graph representation are other WordNet synsets which are related to the synsets present in the text. The arcs of the graph are the relations that hold between these synsets in WordNet. Although, WordNet is a finite structure, a representation of text based on WordNet should only select relevant aspects of WordNet to be included in the text representation. Storylines, therefore, must set the boundaries of the WordNet sub-graph representation while still exploiting the knowledge potential encoded in the relational structure of WordNet as described in section 3.2.2. In order to define the content and delimit the boundaries of this graph representation of an input text, it is necessary to define two criteria:

1. how closely connected to a synset present in the text must a synset derived from text be in order to be included in the text representation. This relates to the notion of semantic distance used in Morris and Hirst's (1991) original lexical chaining algorithm.

2. should certain relations be prioritised? This relates to Hirst and St-Onge's (1998) classification of WordNet relations for lexical chaining.

There are twenty-four binary semantic and lexical relation types that may hold between synsets in WordNet.[12] Some, for example, antonymy or domain membership, are not widely deployed in the database. Others, such as hypernymy, constitute key organising principles of the knowledge base. The different semantics of the relations governing inheritance and transitivity impose some restrictions on acceptable or, at least, meaningful, paths between synsets in the knowledge base.[13] This issue is addressed in Module 3 by means of a classification of relations types into the following three categories:

1. Relations which define an inheritance path to follow
2. Relations to output for all synsets along such an inheritance path
3. Relations to output only if one synset in the binary relation is present in the text

The classification could be made in terms of informativity. Al-Halimi and Kazman (1998) discuss an application that uses a measure of WordNet relation informativeness to prioritise certain relation types. Their rank ordered relations are set out in order of decreasing informativeness in table 5.1. However, this ranking is not empirically justified but is proposed rather as a heuristic to manipulate system performance.

Hirst and St-Onge (1998) propose a classification which is, perhaps, more useful.[14] Semantic relations are classified according to the three directions:

[12] See section 3.2.2 for a full discussion of WordNet relations.

[13] Priss (1998) suggests a formal definition of the synonymy, hyponymy and meronymy relations.

[14] See section 2.6.3 for a review of Hirst and St-Onge's (1998) lexical chaining algorithm.

Rank	Relation type
1	Identity (repetition)
2	Similarity (synonymy)
3	Antonymy
4	Holonymy
5	Hyponymy

Table 5.1: Rank ordering of WordNet relations in (Al-Halimi and Kazman 1998, p.338)

Up, Down and Horizontal, as in table 5.2. According to their definition,

Relation	Direction	Relation	Direction
Also see	Horizontal	Cause	Down
Antonymy	Horizontal	Entailment	Down
Attribute	Horizontal	Holonymy	Down
Pertinence	Horizontal	Hyponymy	Down
Similarity	Horizontal	Hypernymy	Up
		Meronymy	Up

Table 5.2: Classification of WordNet relations into directions adapted from (Hirst and St-Onge 1998, p.308)

Up denotes a generalisation, *Down* a specialisation and *Horizontal* a link such as antonymy which, according to the authors, is "highly indicative of meaning" (p.309). This classification is used to constrain acceptable paths in WordNet for lexical chain construction. Furthermore, they state two rules for inter-synset path construction:

(R1) No other direction may precede an upward link

(R2) At most one change of direction is allowed

(Hirst and St-Onge 1998, p.309)

Although the aim of the Hirst and St-Onge (1998) algorithm is to construct lexical chains for malapropism detection, while the system I developed, Storylines, constructs a graph structure to be used for text comparison, their classification and rules suggest an approach to the problem of constraining the boundaries of the WordNet subgraph under construction.[15] They also provide an interesting point of contrast with the classification defined in this research. The twenty-four relations[16] have been divided according to two classification criteria. The first Storylines classification divides relation types between those which introduce recursion in the system and those which do not. Section 5.6.2 discusses to what extent this correlates with the Hirst-StOnge upward direction class. The second Storylines classification defines relation types as textual or not. This is a somewhat problematic distinction. Section 5.6.2 discusses to what extent this classification correlates with the relations which are defined as lexical versus those which are defined as semantic within WordNet. A textual relation, in this thesis, is defined as a relation which is only relevant for synsets present in the text as opposed to synsets reached by recursion.

[15]It is worth noting that although the Hirst and StOnge algorith in theory is valid for all parts of speech, it was onl implemented for nouns, unlike Storylines which includes nouns, verbs and adjectives.

[16]In the absence of an empirical evaluation of informativity of relation types, it was decided to retain all relation types at this stage of the implementation for processing in Module 4.

Relations for Recursion

The list of relations for recursion are: hypernymy, entailment and causation. Both entailment and causation are defined as downward specialisation relations by Hirst and St Onge. The specification of these two relations in WordNet according to Fellbaum (1998), however, indicates that these two relations are more complex than mere specialisation relations. Indeed, entailment, in a general sense, subsuming entailment, causation and hypernymy for verbs, is the organising principle of verbs within the database (see section 3.2.3). It is not clear where they should stand in relation to the tripartite Hirst-StOnge classification. For the purposes of the Storylines implementation, the chains of synsets set up by these relations constitute vital connectivity within the WordNet verb database which is of equal importance to the hypernymy relation for nouns. The relations for recursion are envisaged as the list of the chain—defining relations in WordNet as a whole. As such, the entailment relations of the verb hierarchy could not be excluded.

The meronymy—holonymy relation, however, was excluded from the list of recursive relations for several reasons. A first justification relates to the Up-Down classification of Hirst and St Onge. In their algorithm, the relationship between a holonym and its meronym is defined as one of specialisation. I, however, do not agree that the relationship between a finger and a hand constitutes a specialisation of meaning. A second formal justification for their exclusion regards the fact that the interaction between hypernymy and meronymy is not clearly defined (see section 3.2.2) and could lead to inconsistencies in the text representation output. A final empirical justification is based on an investigation of meronym chain length and the distribution of meronymy relations through the hierarchy. This analysis indicates that the coding of these relations, like that of the hypernym hierarchy, suffered from

a certain unevenness as well as an over-zealous focus on biological minutiae. Over 70% of nominal synsets are not coded with holonym or meronym relations. The meronym chains that do exist between the remaining 30% of the noun hierarchy are up to fifteen links in length for chains based on a single meronym type and up to seventeen for chains consisting of links of more than one meronym type. Membership meronym chains are up to 12 links in length (e.g. beaver rat — animal kingdom), part meronym chains 15 links in length (e.g. occipital gyrus — brain — organic structure) and substance chains up to 6 links (e.g. linoleic acid — bean curd). The inclusion of meronym chains could only compound the difficulties caused by the imbalance between over-specification and under-specification throughout WordNet. The explosion in the potential number of associated synsets and the unclear formal properties of meronymy in relation to other database relations led to its exclusion from the Storylines list of relations for recursion and its limitation to hypernymy, entailment and causation.

Textual Relations

The textual classification is intended to avoid spurious connections based on lexical relations where the lexical items do not appear in the text. The only relations which are classified as "textual" (i.e. only relevant where one synset is present in the text) are domain membership and derivation relations. The motivation for excluding the domain membership relation is that membership in a domain in WordNet is a question of terminology rather than a conceptual connection. For example, although the terms "set point" and "ace" are in the "tennis" category domain in WordNet, "tennis racket", "tennis player" and "tennis ball" are not. This relation based on a seemingly arbitrary selection of certain terminology for domain membership, should be excluded from the

text representation unless the related terms are actually present in the text. The derivation relations are clearly specific to the word *form* as opposed to word *meaning*, unlike antonymy which, although a relation between word forms, defines a conceptual contrast. All other relation are output at all levels of recursion.

Recursive Relation Extraction Process

Storylines uses the organising principles of WordNet rather than a notion of maximum path length to define which synsets are "associated" with the synsets present in the text and should, therefore, be included in the output representation. There is no constraint on how far apart two synsets may be for the connection between them to be included in the final text representation built in Module 4 (see section 5.7). As, in Hirst and St-Onge's (1998) algorithm, the direction rule set impose criteria for acceptable path traversal in WordNet, so, in the system I developed, Storylines, the list of recursion relations set the boundaries of the relation extraction algorithm. At each level of recursion, the system outputs the relations between the current synset and all synsets at **one** remove from it. This is a more relaxed approach than a fixed criteria of semantic distance and is governed by the organising principles of the knowledge base itself.

Algorithm 5.3 expands on line 11 of the basic algorithm, Algorithm 5.2, outlining the recursive findRelation subroutine. The three lists of relations involved are as follows:

- **recurseRelationList**: hypernymy, entailment and causation
- **alwaysOutputRelationList**: hypernymy, meronymy, holonymy, entailment, causation, similarity, also see, verb group, antonymy and attribute

- **outputTextRelationList**: domain term, derivationally related form, participle of verb, pertainym

The input to the **findRelations** subroutine is a synset and an integer flag with value 1 if the synset is present in the text and 0 if it is not. The first step iterates through the **alwaysOutputRelationsList** array and outputs an appropriate output line, such as example 5.16 for each synset connected to the current synset by this relation.

(5.16) #s 13808178n:13836665n 1 1 0 NI:0 null

If the synset is present in the text ($inTextFlag == 1$), the same process is repeated for the **outputTextRelationList** array. The final step begins the recursion. For each synset connected to the current synset by each relation in the **recurseRelationList** array, the **findRelations** subroutine is called with this synset and the *inText* value 0 as arguments. The final output is a list of relations, as in example 5.16, above which is processed and realised as a graph in Module 4 of the implementation.

Algorithm 5.3 Module 3: Recursive FindRelations SubRoutine

Input: synset *syn*, sentence number *sentence*, flag *inTextFlag*

Output: *relList* List of binary relations between WordNet synsets associated with input *syn*

1: **for all** relationType $R1 \in alwaysOutputRelationList$ for *syn* **do**
2: get $targetSynsetList_R1$
3: **for all** $targetSyn \in targetSynsetList$ **do**
4: **output** $R1$ *syn* $targetSyn$
5: **end for**
6: **end for**
7: **if** $inTextFlag = 1$ **then**
8: **for all** relationType $R2 \in outputTextRelationList_{syn}$ **do**
9: get $list2$ of target synsets for relation $R2$
10: **for all** target synset $S2 \in list2$ **do**
11: **output** $R2$ $S1$ $S2$
12: **end for**
13: **end for**
14: **end if**
15: **for all** relationType $recurseRel \in$ recurseRelationList of $S1$ **do**
16: get $recursionList$ of target synsets for relation $recurseRel$
17: **for all** target synset $recurseS \in recursionList$ **do**
18: **findRelations**($recurseS$, *sentence*, 0)
19: **end for**
20: **end for**

5.7 Module 4: Graph Construction

Algorithm 5.4 Module 4: Graph Building Basic Algorithm

Input: as follows:

1. *relList* list of binary relations (see appendix C.5.4)
2. *sentences* sentence-numbered source text
3. *topologyInfo* WordNet topological features file (see appendix C.5.5)

Output: *WNDag* directed acyclic graph representation of text

1: Load *topologyInfo*

2: Create *WNDag* using *sentences* and *relList*

The final module of the process was implemented as an object-oriented program in Java. It is described in two stages. This section describes the basic algorithm for the construction of the text representation. Section 5.8 outlines the program classes. The procedure is set out in Algorithm 5.4.

The first step of the algorithm involves loading a file derived from WordNet as a static HashMap called synsetHash. This HashMap contains the data for the 111,761 noun, verb and adjective synsets in WordNet which the program requires. Each record in the input file derived from WordNet relates to a unique synset in the knowledge base. The record consists of nine tab-delimited fields and includes values of topological features (see section 4.4) for that synset in the WordNet structure. The record fields are set out in table 5.3 and some sample records in table 5.4. The synonymous word list field, field **9**, itself consists of three fields for each of the synonymous words in the Synset's list of synonymous words, delimited by the percentage (%) symbol:

1. Word form (e.g. "electricity")
2. Word frequency, according to WordNet sense tagged corpus (e.g. 1)
3. Word polysemy count: the number of synsets to which this word belongs (e.g. 2)

For each processed record, a new object of type Node is created and inserted into the synsetHash hashMap. This Node class is described in detail in section 5.8.1.

1	POS	Part of speech
2	Offset	Synset offset: unique identifier for each synset
3	Total	Total number of connections from that synset
4	Sisters	Number of coordinate terms for that Synset
5	Depth	Maximum depth of the Synset in WordNet
6	BF	Synset branching factor (number of children)
7	Leaf	LeafNode: binary field indicating if leaf node
8	Top	TopNode: binary field indicating if top WordNet node
9	Synonymous	Synonymous word list

Table 5.3: Fields of WordNet Topological Input File

POS	Offset	Total	Sisters	Depth	BF	Leaf	Top	Synonymous
n	07054143	2	2	5	0	1	0	electricity%1%2
v	00777368	8	10	5	7	0	0	authorize%32%1,pass%0%42
a	02899542	1	0	0	0	0	0	Cyrillic%0%1

Table 5.4: Sample of WordNet Topological Input File

The second step involves initialising a WNDag object to represent the input text being processed. The WNDag class is the basic text representation

class. It defines a directed acyclic graph where the nodes are synsets and the arcs are relations between those synsets. An arc may include more than one relation type. The program builds a WNDag object for each text. It takes as input the original text file with a single sentence per line (generated in Module 2) and a file of binary relations, such as that set out in table 5.5 and described in section 5.6.1.

Relation	Synset1:Synset2	Sentence	InText1	InText2	term1:index1	term2:index2
pol	00059106n:01173759n	1	1	1	Strike:3	Strike:0
@	01172321n:01173759n	1	0	1	null	Strike:0
+	01173759n:02339744v	1	1	0	Strike:0	null

Table 5.5: Sample binary relations file

The basic WNDag initialisation algorithm is outlined in Algorithm 5.5. The process initialises the three main fields of the WNDag text representation object:

- *text*
- *relationsHash*
- *nodesPresentHash*

These fields are described in section 5.8.5 but the following briefly summarises the initialisation process. A new object of type Text is assigned to the *text* field. This object is created from the input text file which is converted into an array of sentences. The *relationsHash* HashMap object is initialised with an initial capacity equal to approximately the number of lines in the input relations file, as there can only be as many individual relations as there are lines in the file. Similarly, the maximum number of synsets is twice the number of relations, as each relation is between two synsets. Therefore, the

nodesPresentHash HashMap object is initialised to double this size. The *relationsHash* HashMap has key-value pairs of type String and Arc. The String key is the Synset1:Synset2 field from the input relations file, such as the two examples in 5.17

(5.17) 00001740n:00002565n, 05154236v:04562341n

The Arc object value is the relation that holds between these synsets.

For each input line, if there does not exist an Arc object in *relationsHash* for the two listed synsets, a new arc is created and added to *relationsHash*. If there does exist an Arc object between these two synsets, the arc is augmented with the new information in the input line (relation type, sentence number, source term). Likewise, the *nodesPresentHash* has key-values pairs of type String and NodeInGraph. The key String object is, again, the synset offset and its part of speech. For each synset, if a NodeInGraph object representing the synset does not exist in *nodesPresentHash*, a new one is created for this synset using the information from the input line. If such a node already exists, the node is augmented with information from the input line. The final step prunes all leaf nodes that are not present in the text. The only such nodes are nodes related by meronymy, antonymy, derivation or domain to synsets in the text but whose presence was not reinforced by any further connection to any other synsets in the graph.

5.8 Text Representation Data Structures

The graph structure built in Module 4 of the process is a directed acyclic graph containing a subset of the WordNet linked structure augmented with information derived from the source text. Each node in the graph represents

Algorithm 5.5 Module 4: WNDag Graph Construction

Input: as follows:

1. *relList* list of binary relations (see appendix C.5.4)
2. *sentences* sentence-numbered source text
3. *topologyInfo* WordNet topological features file (see appendix C.5.5)

Output: *WNDag* directed acyclic graph representation of text

```
1: Initialise WNDag object: parameters sentences and relList
2: Initialise text field as new Text object: parameter sentences
3: Initialise relationsHash hashMap object
4: Initialise nodesPresentHash hashMap object
5: for all line ∈ relList do
6:    Tokenise line: token1 = arc token2=synset1 token3=synset2
7:    if arc ∈ relationsHash then
8:       modifyArc(line)
9:    else
10:      new Arc(line)
11:      add Arc to relationsHash
12:   end if
13:   for all synset ∈ line do
14:      if synset ∈ nodesPresentHash then
15:         modifyNode(line)
16:      else
17:         new Node(line)
18:         add Node to nodesPresentHash
19:      end if
20:   end for
21: end for
22: Prune all non-text leaf nodes
```

a WordNet synset with added topological (see table 5.3) and text-based features. The intuition behind this representation is that the main entities and topic areas in a text are those linked by frequently traversed edges in the graph. This structure captures the lexical cohesive network structure formed by reiteration, through general nouns, synonymy, superordinates and repetition, by pronominal co-reference (to a limited extent) and by other collocational cohesive devices as edges that have been traversed more than once.

A very simple example of this principle is illustrated by the following short text, *Text1*, based on an extract from (Halliday and Hasan 1976) p. 279.

> There's a boy climbing that tree. The idiot will fall if he's not careful. Elms are not very sturdy. The poor child might hurt himself.

These four sentences yield, among other chains, the lexical chains connecting boy-child-idiot and tree-elm shown in figure 5.2. The width of the line in the figure is directly proportional to the frequency of the particular arc in the input. Different senses of words are written *word#1, word#2, etc.* The text graph also contains the spurious path, related to the lexical item "tree", shown in figure 5.3. While the flora-tree path is reinforced by the addition of the hypernym path for elms, the abstraction-tree path is not reinforced by connections with any other nodes and indeed remains the only node stemming from the root abstraction in the text.

In practice, the implementation does not rely on a simple count of edge traversal, as in this short example. It exploits topological features derived from WordNet, information on word polysemy and frequency and relation type prioritisation to refine this basic count to a more defined measure of

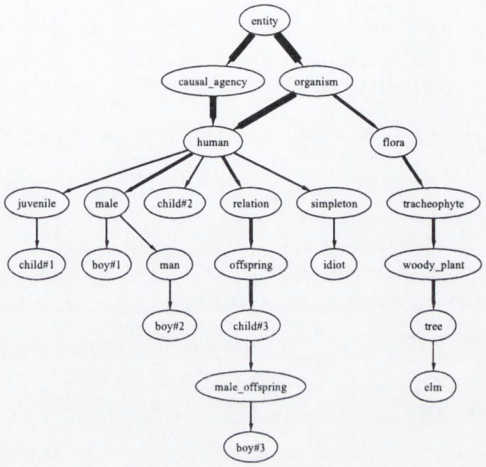


Figure 5.2: Partial hypernym graph for Text1

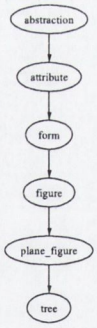


Figure 5.3: Path from hypernym graph for Text1

edge and node weight. Chapter 6 discusses the weighting algorithm in detail with reference to text comparison. This section describes the data structures and classes used to model this directed acyclic graph and how the features used in the weighting algorithm are stored.

The following sections describe the classes representing WordNet synsets (Node.java, § 5.8.1), source text synsets (NodeInGraph.java, § 5.8.2), edges connecting the source text synsets (Arc.java, § 5.8.3), the source text itself (Text.java, § 5.8.4) and the class that synthesises all of the afore-mentioned classes into a unified whole to give a meaningful representation of the source text (WNDag.java, § 5.8.5).

5.8.1 The Node Class: WordNet Synsets

Node objects are designed to represent WordNet Synsets. In WordNet, much of the information regarding synsets is encoded in their relations to other synsets in the knowledge base. Storylines Module 3, the relation extraction process, handles the aspects of this relational structure which are relevant to particular texts and are required by Storylines to build a representation of the text. What remains of WordNet for the system to exploit are features of individual synsets and this is what the Node class represents for individual synsets. The features are listed and described below. The functionality of this class is quite limited, restricted as it is to accessor and mutator methods. This class is essentially a synset data storage class whose fields are used in the calculation of arc weights for text comparison. The fields of the class can be divided into three groups:

1. basic data fields;

2. WordNet topology fields;

3. fields related to the synonymous word list;

The basic data fields are the part of speech of the synset in question and its offset, the eight numeric character unique identifier of the synset in WordNet. The topological fields represent the relative position of the synset within the topology of WordNet. The features used are those which, according to the experiment described in chapter 4, appear to be good predictors of specificity of WordNet synsets. These are as follows:

- **depth** (*int*): the maximum depth of the synset in WordNet, i.e. the longest path from the top node of the hierarchy to the synset in question.
- **branchingFactor** (*int*): the number of children of this synset, i.e. how many synsets inherit directly from this in WordNet.
- **sisters** (*int*): the number of coordinate terms or "sisters" of the synset. This is equal to the branching factor value of the parent synset in WordNet

Related to these topological fields, there are also three topological static constants. The constants store the maximum values for the three fields depth, sisters and branchingFactor. As these values differ for nouns and verbs, the values are stored as an array of two integers. These are as follows:

- **MAXDEPTH** (*int[]*) maximum depth value: noun=16, verb=13
- **MAXSIS** (*int[]*) maximum number of sisters: noun=798, verb=598
- **MAXBF** (*int[]*) maximum branching factor: noun=393, verb=393

The final set of fields relate to the list of synonymous words of which the synset is composed. Storylines exploits particular characteristics of words

that appear in source texts, such as frequency and polysemy, in the calculation of node and arc weights in the representation. The word list related fields, then, are as follows:

- **words** (*java.lang.String[]*): an array of Strings whose values are the synonymous words in the synset list
- **wordFreq** (*int[]*): an array of integers whose values are the frequency of each synonymous word (co-indexed with the words array) from the WordNet sense tagged corpus
- **wordPolysemy** (*int[]*): an array of integers whose values are the number of synsets to which each synonymous word (co-indexed with the words array) belongs

5.8.2 NodeInGraph: text-based information

The NodeInGraph class represents synsets which appear in or are associated with an input text. Each NodeInGraph object is linked to a particular Node object, i.e. a particular synset. The association between Node and NodeInGraph objects is maintained as each NodeInGraph has a unique Node object as a field.[17] NodeInGraph objects represent both fixed semantic and structural information derived from WordNet and dynamic information derived from text. The kind of text-specific information stored is dependent on

[17]Conceptually, a NodeInGraph object is an instantiation of a Node object in the "real world" of a text and this might best be represented as an inheritance relation. However, the static HashMap of Node objects in the WNDag class contains 117,671 unique objects of type Node. Therefore, in order to avoid multiple copies of potentially thousands of Node fields, the relationship between Node and NodeInGraph objects was realised as an object field reference.

whether a NodeInGraph object represents a synset which was present in the text or a synset that is related to one present in the text. For NodeInGraph objects that represent synsets present in the text, the instance fields, set out below in section 5.8.2, record the exact terms used and repetition and synonymy counts. The main functionality of the class, besides representing text synsets, is to calculate the weight of a given NodeInGraph object. The weights, which will be discussed in chapter 6, are calculated on the basis of a combination of fixed WordNet topological features and text-based repetition and synonymy counts defined in the techniques.

Fields

- **node** (*Node*): the node field of type Node refers to a unique node in the WNDag static synsetHash HashMap. No two NodeInGraph objects from the same text, i.e. in the same WNDag object, may refer to the same Node object. Only NodeInGraph objects from different WNDag objects, i.e. derived from different texts, may refer to the same Node object. In summary, each WordNet synset has only a single representation in each text, although this single representation maintains all information regarding multiple instances of a given synset in a source text.

- **inText** (*Boolean*): true for synsets present in the text, false for those derived from it

- **numberOfTermsPresent** (*int*): the number of times a term from this synset has appeared in the text. This amounts to a count of all repetition and paraphrase in a given text

- **int numberOfDifferentTerms** (*int*): the number of different word

(or multiword) forms from this synset that have appeared in the text. This amounts to a count of how many different synonyms have been used in the text but not how many times each synonym has been used.

- **termsPresent** (*int[]*): this array is co-indexed with the **words** array of the **node** instance field. The values of this array are a count, for each synonymous term in the synset, of how many times this term was used in the text.

- **inSentences** (*int[][]*): this two-dimensional array is also co-indexed with the **words** array of the **node** instance field and also with the **sentences** array of the Text object, i.e. the sentences of the source text. Its values can be summarised as:

  inSentences[x][y] = count of word at index x in the **words** array (the synonymous word list) in sentence y

5.8.3 Arc: WordNet and Text relations

The Arc class models the arc between two NodeInGraph objects of the WNDag directed acyclic graph. An Arc object maintains the list of relations that hold between two synsets and a count of edge traversal for each relation type. The possible relation types are listed at section 5.6.2. In practice, excluding polysemy relations, it is not common for Arc objects to include more than one *type* of relation, such as antonymy or meronymy, but it is theoretically a possibility. This class represents the relational structure of WordNet which Storylines aims to exploit in representing text content and estimating text similarity.

Fields

The fields of the class may be divided into three categories: references to the source and target nodes, fields to record relation types and relation counts and fields to record Arc weights.

1. Source and target nodes:

   - **nodes** (*NodeInGraph[]*): an array containing references to the Arc object's source and target NodeInGraph objects.

2. Relation fields

   - **relations** (*int[]*): Array of relation counts. Each index refers to a different type of pointer. The array values are counts of occurrences for each type of pointer.
   - **sentencesIn** (*int[][]*): Two-dimensional matrix where rows refers to relation types and columns to sentences in the source text.

     sentencesIn[x][y]=count of the times relation type x occurs in sentence y

3. Weights

   - **listOfArcValuesForComparison** (*double[]*): array of weights. The indices of the array are techniques for the different methods of weight calculation which are discussed in section 6.6.2.

Behaviour

The main functionality of this class is Arc weight calculation. These calculations are based on several criteria:

- count of arc traversal, mitigated by polysemy count of NodeInGraph objects or not.
- presence in text
- source and target NodeInGraph object weight values
- relation type

Subsequent sections discuss how these criteria contribute to a measure of information content in the graph (§ 5.9) and how different combinations of these are employed to determine graph similarity or difference (§ 6.6.2).

5.8.4 Text: the original text

The Text class is simply a data storage class for the source text. The only field consists of an array of sentences.[18] This array is co-indexed with the *inSentences* and *sentencesIn* fields of the NodeInGraph and Arc classes respectively. Its purpose is to store the source text sentences for output if deemed of significant difference between two similar texts (see section 6.8)

5.8.5 WNDag: the text representation class

The WNDag class constitutes the meaningful text representation which Storylines was designed to build. It is intended to capture the lexical cohesive structure formed by repetition, paraphrase and collocational relations in the source text. The structure derived is a labelled directed acyclic graph which consists of NodeInGraph objects for nodes and Arc objects for edges. The

[18]The list of sentences is derived from the LT CHUNK tagged output in Module 1 (see section 5.4). Sentence boundaries are marked with a <S> tag. The co-indexing of words with sentences derives from the processing of the LT CHUNK output at Module 2.

richness of the structure's information content is distributed through these Arc and NodeInGraph objects. They represent the semantic content of the WordNet database relevant to a particular text, the relational structure of WordNet relevant to that text, the text-based cohesive relations in that text and they store structural information about WordNet by means of which its semantic content may be evaluated. The interconnections of which a WNDag object is composed reflect the network structure of the text which it is intended to represent.

Instance variables

In practice, the graph structure is represented by two HashMap objects:

1. **nodesPresentHash**: a HashMap of all NodeInGraph objects derived from the source text. The HashMap keys are String objects consisting of synset offset keys and the synset part of speech, as in example 5.18.

   (5.18) 00001740n, 05154236v

   As such, there is a unique key for each key-value pair. This HashMap stores the graph nodes.

2. **relationsHash**: a HashMap of all Arc objects derived from the source text. The HashMap keys are String objects consisting of the HashMap key of the source and target node, connected with a colon:

   (5.19) 00001740n:00002565n, 05154236v:04562341n

   The key is unique for each key-value pair. This HashMap encodes the edges of the graph.

The use of the *java.util.HashMap* ensures constant time performance for basic access operations for nodes and edges of the graph. Section 6.3 discusses how

this addresses the potential intractability of the NP-complete graph matching problem inherent in the task of determining graph similarity.

Static variables

In addition to the two instance variable HashMap objects, the WNDag class stores the static HashMap of Node objects to which all NodeInGraph objects refer.

1. **synsetHash**: a HashMap of Node objects loaded from pre-compiled WordNet data file. The *synsetHash* and *nodesPresentHash* keys are identical. There are 117,671 unique Node objects loaded for each of the 117,671 unique WordNet synsets. The first step in any application using the WNDag class is to load the WordNet data file and populate the **synsetHash** HashMap.

Behaviour

The behaviour implemented for this class is largely related to the task of similarity identification and differential information detection and to the task of visually representing text. The visual representations are designed to be rendered using AT&T's graphviz graph plotting program (Gasner *et al.* 1993). The output is colour-coded for presence in text and arc type and uses arc weight to determine edge width in the output graph. The similarity estimation and differential content detection implementations are discussed in detail in chapter 6.

To a certain extent, this discussion of implemented behaviour belies the real potential of the class. The current implementation is focused on the similarity detection task. However, the real strength of this graph representation lies in the latent information it contains which could be exploited for

any number of other tasks. The class encapsulates the information derived from the form and content of the WordNet knowledge base as well as from the source text in a form which allows access and search in constant time and traversal in linear time. The information content of the graph is discussed in greater detail in section 5.9. The possible tasks which could exploit this structure are numerous, the most viable of which are discussed in the future work section 9.2.

5.9 Latent information content

Section 5.8 set out how and where information is stored in the Storylines text representation model, the WNDag object representing a source text. This section discusses the potential of this stored information in estimating the information content of the model or parts thereof. The types of stored information can be divided into the three categories set out in table 5.6 Category I

Category	Type of information
Category I	text-based term usage information
Category II	conceptual information from a knowledge base
Category III	knowledge base topological information

Table 5.6: Categories of Model Information Types

provides repetition and paraphrase counts for individual synsets where both repetition and paraphrase are identified as key techniques to build textual cohesion. Category II comprises the semantic content of terms in a text and the relationships that hold between different terms. This is the contribution of the knowledge base whereby individual terms are supplemented with information about concepts with which they are associated. This in-

formation builds a representation of at least some aspects of collocational cohesive devices in a source text. Indeed, the more types of connectivity in the knowledge base, the more aspects of collocation could be represented. As was discussed in section 2.5, the structure built by lexical cohesive devices, whether in the lexical chains of Halliday (Halliday and Hasan 1976) or the lexical nets of Hoey (Hoey 1991), can provide an accurate and useful model of text meaning. This theory has been tried and tested in numerous applications of lexical cohesive devices to NLP tasks (see section 2.6.4), often making use of Category I and II type information.

One of the main aims of this implementation was to avoid the traditional pitfalls of employing Category II information from a hand-coded knowledge base such as WordNet. In this aim, the research follows in the tradition of Resnik's information content measure (Resnik 1999). Resnik's model addresses the idiosyncrasies of taxonomic structure by adding a corpus-derived measure of information content to taxonomy concepts. In the Storylines model discussed here, the question of knowledge base population density or sparseness is addressed by features derived from the shape of the knowledge base itself, Category III above. These features were selected on the basis of the results of the concept specificity experiment discussed in section 4.3. The depth, branching factor and sister features appeared to be the best predictors of concept specificity in the knowledge base. Storylines makes the assumption that comparable specificity entails comparable information content. Hence, these topology measures provide a means of factoring out the non-uniformity of distances between concepts in WordNet and of evaluating the individual contribution of nodes to the information content of the whole.

In summary, Category I and II features build a structure which models several aspects of the lexical cohesive nature of text and, hence, the semantic

content of text. The combination of these two types with Category III features produces a model wherein the effects of the choice of knowledge base have been normalised to a certain degree. This structure can be more accurately pruned or rendered pictorially or compared with other structures of the same type. Chapter 6 goes on to discuss how this weighted representation is employed in the task of text comparison.

5.10 Discussion

This chapter describes the text representation component of the Storylines system I designed and implemented. The development of this system addresses several of the research aims of this thesis.

1. Storylines builds a representation of text from free text without mark-up or disambiguation.

2. It includes all open class parts-of-speech (excluding adverbs) in the representation built.

3. Storylines captures aspects of the lexical cohesive structure of text in the text-based information and WordNet semantic content stored in the graph representation of the source text.

4. It provides the facility to manipulate or control how the WordNet semantic content information is interpreted through the use of topological features of the knowledge base which were identified in an empirical analysis of concept specificity.

5. The representation built could be used in many other NLP applications as the HashMap data structures used to implement the design perform access and traversal operations in constant and linear time respectively.

Chapter 6 discusses how the Storylines representation described in this chapter has been applied to the task of similarity estimation and differential sentence extraction of texts. Chapter 8 evaluates the similarity detection performance of the system on two corpora of news texts.

Chapter 6

Text Comparison: Identifying Similarities and Differences

6.1 Introduction

As John Locke wrote in the seventeenth century:

> No knowledge without discernment.

In the 21st Century, the average human needs a high capacity for discernment to get any knowledge from the huge quantities of information available to them. This chapter addresses the issue of discernment among text documents which has been formulated as the question of distinguishing texts from each other. Discernment, as it is proposed in this chapter, is concerned with similarities and differences. Given a collection of documents, it is useful to determine which documents are similar to each other. It is even more useful to add where the documents differ. This kind of precise content feedback is a very useful tool to help users digest information content from multiple documents.

In his seminal article on similarity, Tversky (1977) highlights the importance of similarity in human cognitive processes.

> Similarity plays a fundamental role in theories of knowledge and behaviour. It serves as an organizing principle by which individuals classify objects, form concepts, and make generalizations. (Tversky 1977, p.327)

As a result, a central problem for systems that aim to model or emulate human processes involves measuring similarity. In the information sciences, this problem arises in several domains. This thesis is concerned with semantic similarity between text documents, a core area of the Information Management and Retrieval domain. Chapter 5 detailed a structure for meaningful text representation. This chapter discusses how this structure has been employed to determine similarities and distinguish differences of content in texts,

independent of its surface realization. According to Gentner and Markman (1994):

> the process that determines similarity may be profitably characterized as alignment and mapping between structured representations. (Gentner and Markman 1994, p.157)

This chapter details the alignment of similarities between two text representations under comparison and the evaluation of the informational contribution of these similarities to the semantic content of the texts, both separately and together. Having determined similarity, the complementary task of discerning differences in the aligned representations is tackled.

This alignment and discernment system has been fully implemented in Java and details of the design issues and decisions involved are set out in subsequent sections. The specific aims of this implementation are set out in section 6.2. Section 6.3 discusses the complexity of the graph-matching problem. Section 6.4 outlines the basic algorithm. Similarity and difference graph derivation (see § 6.5) and measurement (see § 6.6) are discussed in detail. Section 6.6.1 describes model features employed individually and in combination by the measurement and weighting algorithms according to particular technique formulae set out in section 6.6.2. Section 6.7 outlines the similarity coeffients used to determine significance of the information overlap. Section 6.8 goes on to discuss the issue of text difference and how key differences between texts are selected. Finally, section 6.9 discusses the contributions of this research. A thorough performance evaluation is discussed in chapters 7 and 8.

6.2 Aims

Storylines, the system reported in this dissertation, was implemented with two primary aims in mind. The first aim was to exploit the potential of the meaningful representation of text described in chapter 5 in the task of similarity and difference detection. As I noted in the introduction, determining document to corpus or document to document similarity is a core task in the information sciences in general and Natural Language Processing in particular. Storylines, the text representation system implemented during the course of this research was designed to exploit the WordNet knowledge base for its semantic information while controlling how this information might be interpreted based on a weighting scheme derived from the structure of WordNet itself. Simpler structures have proved successful to varying degrees on the task of similarity detection. Indeed, the implementers of the Google News website, which groups news texts from different sources by similar content, by-pass text representation completely and similarity detection is carried out "entirely by a mathematical algorithm, based on many factors including how often and where a story appears elsewhere on the web".[1] The intention of this implementation is to determine whether the performance of the system justifies the extra computation involved in building a richer representation of text.

The second and related main aim of this implementation is to design a tool to pre-filter texts by giving feedback on similarities and differences between text pairs. Such a tool would allow users to make an informed choice about the information content of texts and to determine, given a selection of texts, which text is most worth reading. A current "real world" application for this tool is the Google News website. The selection of news texts provided

[1]From personal correspondence with the Google News team.

by Google give an overview of how a topic is treated in the world's English language media. What is lacking, however, is a further content analysis of these similar stories to highlight their content differences. Based on this, the user could make a choice as to which text is the most informative to read. Section 6.8 details the component implemented to carry out this task. Chapter 8 examines the system's performance on both the similarity and difference detection tasks.

6.3 Complexity of the problem

In the introduction 6.1, the task of similarity detection was proposed as one of "alignment and mapping between structured representations" (Gentner and Markman 1994, p.157). The structured representations in this implementation are directed acyclic graphs. The alignment task, therefore, could be viewed as a graph-matching problem. According to Garey and Johnson (1979), graph isomorphism is an open-problem. Skiena (1997) goes on to state that:

> No polynomial-time algorithm is known for graph isomorphism, but neither is it known to be NP-complete

However, as the graphs to be compared may be of different sizes, the problem could be interpreted as a sub–graph matching problem which is known to be NP-complete. The potential intractability of the graph– or subgraph–matching problem could prove an insurmountable stumbling block to this implementation. However, as with many programs where the theoretical possibility of intractability looms, in practice, the input is clearly bounded and hence, the spectre of exponential time complexity does not become a reality. For the implementation described here, the WordNet graph from

which all of the text sub-graphs derive sets the upper bound for any input to the program. Theoretically, the text data structures are finite with a maximum size equal to that of WordNet at 117,561 nodes and 333,827 edges. In fact, the average number of edges for texts in the test corpora is only approximately 10,000. Furthermore, the problem is re–formulated in this implementation as a simpler task of edge–matching where each input graph is represented as an associative array of edges. The complexity of this algorithm as it is implemented here is linear on the number of edges, requiring a single traversal of the associative array. Therefore, although, in theory, the problem of graph– or subgraph–matching is probably intractable, in practice, this implementation simplifies the problem to one which is not just tractable but fully practical.

6.4 Basic algorithm

The procedure for the similarity detection and difference identification task is set out in Algorithm 6.1. This component of the Storylines system takes as input two WNDag objects generated from free text, as described in chapter 5. It outputs the result of the similarity measure for the two texts. If the two texts are judged to be similar, it outputs the sentences which have been identified as the key information differential between the two texts. There are four main procedures involved in this task:

1. Derivation of the WNDag objects of the intersection of the two input graphs and the two graphs minus the intersection (see section 6.5);

2. Calculation of the size of each of the graphs according to the weighting technique selected (see section 6.6);

Algorithm 6.1 Similarity Detection and Difference Identification

Input: Two graphs $G_1 = (V_1, E_1), G_2 = (V_2, E_2)$ where V_1 is the set of vertices and E_1 is the set of edges of G_1 and V_2 and E_2 are the set of vertices and edges, respectively, of G_2.

Output: as follows:

1. Similarity judgment
2. 2 most different sentences (if similar)

Derive three graphs (WNDag objects):

1. $G_{common} = (V_{common}, E_{common})$ representing $G_1 \cap G_2$
2. $G_{1less2} = (V_{1less2}, E_{1less2})$ representing $G_1 \backslash G_2$
3. $G_{2less1} = (V_{2less1}, E_{2less1})$ representing $G_2 \backslash G_1$

Calculate size of G_{common}, G_{1less2} and G_{2less1} according to chosen technique

Evaluate size of G_{common} wrt to G_{1less2} and G_{2less1}

Compute significance of size

if Size of G_{common} is significant **then**

    Output "is similar"

    Identify and output sentences with most differential content in G_{1less2} wrt G_{common}

    Identify and output sentences with most differential content in G_{2less1} wrt G_{common}

else

    Output "is not similar"

end if

3. Evaluation of the size of the intersection graph relative to the size of the complement graphs (see section 6.7);

4. Selection of key differential sentences (see section 6.8);

There are two questions of program efficiency that should be mentioned at this point. Firstly, the four procedures of the algorithm are presented separately in this discussion for modular clarity. In practice, however, graph derivation, size calculation and differential sentence selection can occur simultaneously. Each individual arc contributes to the calculation of overall graph size and to the selection of differential sentences independently. Therefore, as each new arc is added to one of the intersection or remainder graphs, its contribution to graph size and sentence count may be incorporated. The final process, when all arcs are allocated to one of the three graphs, is the evaluation of intersection size relative to the two remainders and, if this is deemed significant,[2] the differential sentences are printed. This approach entails traversing the two source WNDag objects without additional traversals of the three derived graphs and is, of course, much more efficient. However, for clarity, rather than conflating the three processes in one, they are presented separately below.

The second question of efficiency is one of memory allocation. As it is set out here, the similarity detection algorithm takes as input two WNDag objects representing two texts. It then goes on to create three more WNDag objects to represent the intersection of the two texts, the first text minus the intersection and the second text minus the intersection. It is not a requirement of the approach that these three WNDag objects are built. However, it is useful to examine and render these graphs pictorially. For that reason, it was decided to assign objects for each. This design decision, however, can

[2]The significance calculation is described in section 6.7.

easily be reversed without altering the overall idea that I have designed and tested for this dissertation.

6.5 Deriving the intersection and difference graphs

The procedure to separate the common and the different elements of two input WNDag objects is set out in Algorithm 6.2. Constructing the intersection of the two WNDag objects (e.g. $Text1$ and $Text2$) involves two tasks:

1. Determining what the two WNDag objects have in common
2. Merging their common information

Identifying commonalities is accomplished by iterating through all Arcs of $Text1$ and testing whether each Arc is present in $Text2$. All WNDag Arc objects are stored in a HashMap object, *relationsHash*. For each key in *relationsHash* of $Text1$, the system tests if that key has a value in *relationsHash* of $Text2$. This test, therefore, can be carried out in constant time once the hash table is constructed. As stated in section 5.7, the *relationsHash* keys are Strings consisting of the synset offset and part of speech of the source synset and those of the target synset separated by a colon as in example 6.1.

(6.1) 00568973n:00985945v

(6.2) 00006789n:01234709n

Arc matching, therefore, is not a function of any features of the Arc but only of the presence of an Arc of any type connecting the two unique synsets in both WNDag objects.

Algorithm 6.2 Similarity and Difference Graph Building

Input: Two graphs $G_1 = (V_1, E_1), G_2 = (V_2, E_2)$ where V_1 is the set of vertices and E_1 is the set of edges of G_1 and V_2 and E_2 are the set of vertices and edges, respectively, of G_2.

Output: Three output graphs $G_{common} = (V_{common}, E_{common})$, $G_{1less2} = (V_{1less2}, E_{1less2})$, $G_{2less1} = (V_{2less1}, E_{2less1})$

1: **for all** e_1 such that $e_1 \in E_1$ and v_j, v_k are vertices of e_1 **do**
2: **if** there is some e_2 such that v_j, v_k are vertices of e_2 and $e_2 \in E_2$ **then**
3: create new NodeInGraph objects n_1 and n_2 from merged vertices of e_1 and e_2
4: create CommonArc object c_i from $e_1 \bigcup e_2$
5: add c_i to E_{common}
6: add n_1, n_2 to V_{common}
7: **else**
8: add e_i to E_{1less2}
9: add v_j, v_k to V_{1less2}
10: **end if**
11: **end for**
12: **for all** e_i such that $e_i \in E_1$ and v_j, v_k are vertices of e_i **do**
13: **if** $e_i\neg \in E_2$ **then**
14: add e_i to E_{2less1}
15: add v_j, v_k to V_{2less1}
16: **end if**
17: **end for**

If an Arc is present in only one WNDag object, the Arc and the source and target NodeInGraph objects are copied to the appropriate difference WNDag object. If, on the other hand, an Arc is present in both WNDag objects, i.e. Arc matching has succeeded, a new Arc is created which unifies the features of the original Arcs and NodeInGraph objects and adds these new objects to the intersection WNDag object.

The CommonArc class represents these unified Arcs. CommonArc objects inherit directly from the Arc class and have all the same attributes, as listed in section 5.8.3. CommonArc objects, however, have an extra attribute to maintain source text sentence information.

- **jointInSentences**: this field is a three dimensional array which amalgamates the **inSentences** two dimensional arrays of the two source Arc classes.

  jointInSentences[x][y][z] = count of occurrences of relation z in sentence y in text x

The Common Arc constructor merges the **relations** arrays of the two matching Arc objects. The values of the array are joint counts for each relation type. Likewise, the fields of the new common NodeInGraph objects, set out in table 6.1, are joint counts of the original NodeInGraph instance fields.[3]

numberOfInstancesInText (*int*)	termsPresent (*int[]*)
numberOfDifferentTerms (*int*)	numberOfTermsPresent (*int*)

Table 6.1: Merged NodeInGraph fields

This step produces a WNDag object of CommonArc objects and merged NodeInGraph objects to represent the intersection of the WNDag objects de-

[3]See section 5.8.2 for a description of the fields.

rived from two source texts and two WNDag objects consisting of those Arcs from the respective original WNDag objects which are not in the intersection and the NodeInGraph objects associated with these Arcs. Section 6.6 is concerned with determining the size of these three WNDag objects in order to subsequently evaluate their individual contributions to the information content of the whole. Again, more compact representations could be constructed but the intent here is to provide an explicit specification of the proof of concept for the similarity estimation task.

6.6 Measuring graph size

The basic procedure used in the Storylines system for measuring the size of text intersection and difference, given these representations, is set out in Algorithm 6.3. The system traverses each of the three WNDag graph objects, derived according to the algorithm described in section 6.5, to calculate an individual weight value for each graph. For each Arc of a given graph, a weight is calculated. The final graph weight is the sum of all its Arc weights.[4] The key point to note here is that the weight is calculated according to some technique p. There are seventeen possible techniques for weight calculation. The list of techniques is specified as a constant of the WNDag class. The choice of technique indicates which of the text-based or WordNet derived semantic or topological features discussed, in section 5.9, to use and in what combination in order to calculate weights for NodeInGraph and Arc objects. Section 6.6.1 sets out the features for weight calculation. Section 6.6.2 goes on to discuss how these are combined for each system technique option.

[4]In practice, the Arc stores a list of weights for all techniques, calculated as each Arc is created or modified. Therefore, the weight calculation is carried out during the WNDag object construction.

Algorithm 6.3 Measuring similarity and difference

Input: weight calculation technique, **p**

Input: Three graphs G_x, where V_x is the set of vertices and E_x is the set of edges of G_x

Output: *areSimilar* Similarity coefficient

1. $G_{common} = (V_{common}, E_{common})$
2. $G_{1less2} = (V_{1less2}, E_{1less2})$
3. $G_{2less1} = (V_{2less1}, E_{2less1})$

1: instantiate three counters to 0: $Weight_{common}, Weight_{1less2}, Weight_{2less1}$
2: **for** $G_i = G_{common}, G_{1less2}, G_{2less1}$ **do**
3: **for all** e_i such that $e_i \in E_i$ **do**
4: calculate weight of e_i according to parameter **p**
5: add to $Weight_i$
6: **end for**
7: **end for**
8: Apply similarity measure to $Weight_{common}, Weight_{1less2}, Weight_{2less1}$ to give coefficient *areSimilar*
9: Output *areSimilar* coefficient

6.6.1 Features for weight calculation

Section 5.9 discussed the three categories of information encoded in the text representation constructed by the Storylines system. This section investigates how the three categories are used to give a measure of WNDag object weight. The Arc, NodeInGraph and Node class fields, discussed in section 5.8, store different category information. In practice, the various class fields are distilled to eight features, listed in table 6.2. These features are used singly or combined in the weighting algorithm to calculate Arc object weight, according to the weighting technique selected. The derivation of these weighting techniques relative to features of the text representation constitutes one of the contributions of this research.

	Abbreviation	Feature	Class derived from
1	**Frq**	traversal count	Arc
2	**Txt**	simple in text count	NodeInGraph
3	**Rep**	repetition	NodeInGraph
4	**Par**	paraphrase	NodeInGraph
5	**Pol**	polysemy count	NodeInGraph
6	**Rel**	relation type	Arc
7	**Spc**	Node specificity value	Node
8	**NSpc**	Normalised node specificity	Node

Table 6.2: Features for Weighting Algorithm

Traversal count

The traversal count feature is simply a count of traversal frequency for a particular Arc object. The intuition is that the greater the frequency, the greater

the contribution of that Arc to the overall information content of the graph. Therefore, frequently traversed arcs constitute important content areas of a text. However, as the WNDag objects are constructed using complete hypernym, causation and entailment chains from source word in a text to the top of the WordNet hierarchy, the more generic concepts towards the top of the hierarchy, for example "entity", "artefact" or "change", will occur much more frequently in the WNDag objects than more specific concepts lower in the hierarchy, for example "fireman", "hose" or "crumple". Furthermore, these more generic concepts at the top of many hypernym and entailment chains are more likely to be found among those concepts common to two texts. Hence, the frequency of generic concepts and particularly their prominence in intersection WNDag objects means that a simple frequency count alone to measure the individual contribution of intersection and difference graphs to the semantic content of the whole is misleading.

Presence in text

Features related to presence in the source text indicate whether a particular NodeInGraph object originates directly from a word present in the text or simply derives from it. While the extra information provided by WordNet is of vital importance to this application, one must not forget the pre-eminence of the source text. Therefore, NodeInGraph objects which originate in the text itself are prioritised, i.e. assigned a greater weight. It should be noted that not all *inText* NodeInGraph objects are leaf nodes in the WNDag data structure. This feature is used to counterbalance the weightings of NodeInGraph objects according to their genericity or specificity as it is important not to underestimate the contribution of internal nodes, perhaps those representing more generic concepts, where they originate in the text. In practice,

presence in text is represented in three ways.

1. **Simple count**

   The simple count gives the number of occurrences of all terms from the synonymous word list of the synset which a NodeInGraph object represents. The count includes singular and plural forms of each term that occurs, independently and without stemming. It does not, however, indicate how many repetitions or how many paraphrases this includes. For example, the synset in example 6.3 represents the daffodil flower.

   (6.3) 11673208 daffodil, Narcissus pseudonarcissus – (any of numerous varieties of Narcissus plants having showy often yellow flowers with a trumpet-shaped central crown)

   If the word "daffodil" appears three times in a text on gardening, "daffodils" twice and the genus name "Narcissus pseudonarcissus" appears once, the simple count for this synset is six.

2. **Repetition**

   As discussed in section 2.5, repetition is a key device in constructing cohesive text. The *termsPresent* array of the NodeInGraph class logs the number of occurrences in the source text of each synonymous word included in a WordNet synset. The repetition feature is a count of the sum of repetitions of each word in the synonymous word list. The count includes only repetitions of terms or plurals of terms, not paraphrase. For example 6.3 given above, ($daffodil \times 3$, $daffodils \times 2$, $Narcissus\ pseudonarcissun \times 1$), the repetition count for this NodeInGraph object is four as the repetition count is incremented only for each *additional* use of any of the words in the synonymous word list.

If, however, "daffodil" appears five times and the genus name twice, the repetition count is five.

3. **Paraphrase**

   Like repetition, paraphrase is a powerful cohesive device. The paraphrase feature for a NodeInGraph object is a count of how many different terms from the synonymous word list occurred in a source text. With reference to example 6.3 again, where daffodil appears five times and the genus name twice, the paraphrase count is two.

A further example illustrates the relationship between these three counts. The WordNet synset (example 6.4), which represents a car has five words in its synonymous wordlist. In the short excerpt below, the term "auto" appears three times, "automobile" once and "car" once. For a NodeInGraph object representing this synset in this text, the paraphrase count is three, the repetition count is also three and the simple count is six. In fact, the simple count is $repetitionCount + paraphraseCount$.

(6.4) 1. (598) 02853224 car, auto, automobile, machine, motorcar – (4-wheeled motor vehicle; usually propelled by an internal combustion engine; "he needs a car to get to work")

Auto racing began almost immediately after the construction of the first successful petrol-fuelled *autos*. In 1894, the first contest was organised by Paris magazine Le Petit Journal, a reliability test to determine best performance.

A year later the first real race was staged, from Paris, France to Bordeaux, France. First over the line was Emile Levassor but he was disqualified because his *car* was not a required four-seater.

With *auto* construction and racing dominated by France, the French *automobile* club ACF staged a number of major international races, usually from or to Paris, connecting with another major city in Europe or France.

(from http://www.wordiq.com/definition/Auto_racing 1/9/04)

The "inText" counts for the source and target NodeInGraph objects of a particular Arc object have a multiplicative effect on the weight assigned to that Arc. Where paraphrase and repetition counts are used together, their combined weight is $paraphraseCount \times repetitionCount$. In lexical chaining algorithms, repetition is treated as the strongest link and paraphrase the second strongest. It is appropriate, therefore, that the combined effect of the two features is their product rather than their sum (the simple count value), because arithmetically this provides a stronger weighting for the two features than a simpler metric.

Polysemy count

The features discussed above derive from counts of instances of words in the source text. However, the source texts are not disambiguated and indeed the final representations include many unresolved lexical ambiguities. In order to avoid over-estimating the contribution of single polysemous words, it may be necessary to normalise the frequency, repetition and paraphrase counts discussed above for polysemy. The system discussed here only factors out polysemy for those NodeInGraph objects which represent polysemous words in the text, not those which derive from polysemous words in the text. This design decision was made for two reasons. Firstly, every synset whose synonymous word list contains a polysemous word that occurs in a source text will be added to the WNDag representation as a NodeInGraph object. If a synset is derived from the text, and its synonymous word list happens to contain a polysemous word, it does not follow that all other synsets which contain that polysemous word will also be added. Therefore, it is only polysemous words in the text which introduce multiple synsets

to the representation, the effects of which must be contained. Secondly, although polysemous words in fact introduce *chains* of synsets rather than individual ones, it is only the NodeInGraph objects representing the word in the text which benefit from the multiplicative effect of repetition, paraphrase and frequency counts and, therefore, mainly these in particular which need to be controlled.

A further point to note is that the polysemy count for each NodeInGraph object is calculated only on those terms which appear in the text, not the total polysemy count for all words in the synonymous word list. For example, the word "nozzle" has two senses in WordNet, listed in example 6.5.

(6.5) 1. 03689365 nozzle, nose – (a projecting spout from which a fluid is discharged)

2. 05278748 beak, honker, hooter, nozzle, snoot, snout, schnozzle, schnoz – (informal terms for the nose)

If the word "nozzle" appears in a text, the polysemy count for synset "03689365" is two, disregarding the thirteen senses of "nose". Conversely, if the word "nose" appears in a text, the polysemy count for synset "03689365" is thirteen, disregarding the two senses of "nozzle". If both words are present in the text, the polysemy count is fifteen.

The combination of polysemy count with any of the *inText* count feature listed in § 6.6.1 is also calculated for each term present in the text. The repetition or paraphrase count for each term is divided by the polysemy count for this term. The $\frac{Txt}{Pol}$ and $\frac{Rep*Par}{Pol}$ terms of the formulae in table 6.7 are calculated in this way.

Relation type

Section 5.6.2 set out a classification for WordNet relations by which the graph structure of WordNet is bounded to give a subgraph representing a source text. The classification distinguishes three categories of relations:

1. those which introduce recursion in the relation extraction process (see section 5.6) and extend the scope of the subgraph
2. those which appear in output at all levels of the derived WordNet graph
3. those which appear in output only for synsets in the text itself.

This tri-partite classification is an implicit rank ordering of relations according to their importance in the relational structure of WordNet and the text subgraph derived from it. Lexical chaining algorithms use relation type to determine the strength of a link in a hypernym chain. The relation types derived from WordNet, however, are often all ranked at the same level behind repetition and paraphrase. Hirst and St-Onge's (1998) *strong* relations, which include all WordNet relations has been widely used for this purpose (Barzilay and Elhadad 1997, Stokes 2004, Silber and McCoy 2002).

In Storylines, all WordNet relations are not treated as equal. The basic distinction is between those relations which constitute key organising principles of the database and those which denote a more amorphous relation, such as "verb group" or "also see" whose meaning and, therefore, contribution to overall content is somewhat unclear. The weights are set out in table 6.3. The highest values are assigned to the core WordNet semantic relations, followed by less central semantic relations, word-form derivation and domain relations. The domain relations have been given quite a low weight as, if a domain is truly relevant to a text, there will be multiple domain relations

in the associated WNDag object and the sum of all these Arcs will be substantial. If a domain is not relevant, it is not desirable to emphasise single Arcs of type Domain. The low weight is assigned to allow only a significant contribution from domain Arcs when there are many of them. The weights are assigned with values between 0 and 1 on the basis of my experimentation and practice in the literature.

The **relations** array of the Arc class maintains a count of the occurrence of each relation type for a particular Arc object. The relation type weighting feature is calculated as the sum of the products of each relation type count by its weighting, according to formula 6.6.

$$\sum_{i=0}^{numberOfRels} (count_i * weight_i) \tag{6.6}$$

Node specificity value

Throughout this thesis, I have discussed the difficulty of using a hand-coded knowledge base whose internal structure is not homogeneous and whose idiosyncrasies are not quantified. Section 5.8.1 described the Node class which stores the topological features of WordNet synsets, i.e. the position of the synset relative to the rest of WordNet. This section sets out two measures of node specificity derived from these topological fields. The purpose of this node specificity measure is to factor out issues of population sparseness or density in WordNet by evaluating the contribution of each Node object relative to its depth in the hierarchy, its connectivity and its siblings. The first measure is normalised for maximum values of the topological features used, the second is not.

According to the results of the specificity experiment, set out in section 4.3, specificity of concepts which synsets denote is correlated with depth

Relation Type	Weight
Hypernymy	0.9
Meronymy	0.9
Entail	0.9
Cause	0.9
Antonymy	0.9
Attribute	0.9
Similar	0.7
Also see	0.7
Verb group	0.7
Related	0.7
Participle of verb	0.5
Pertains to	0.5
Derived from	0.5
Category domain	0.3
Region domain	0.3
Usage domain	0.3

Table 6.3: Relation Type Weights

in the hierarchy and number of sisters. A specificity measure is intended to give a high value for nodes deep in the hierarchy, with many sisters and a low value for nodes high in the hierarchy with few sisters. The experiment did not reveal a significant effect for the branching factor feature on specificity judgments. This was somewhat counter to expectations. Therefore, although not justified by the empirical evidence, the branching factor feature was used to determine node specificity in experiments with the system and proved effective. The formulae set out here are the result of several tests with different weighting measures and this combination appeared to give the best results on the text comparison task and on analysis of the weights assigned to individual synsets in the knowledge base. As noted in section 4.3, it would be useful to carry out a much larger experiment to determine with greater certainty those features of WordNet topology that are the best predictors of concept specificity. In the interim, those features which resulted most informative in the completed experiment will be exploited.

The formulae for the two measures I propose, Spc and $NSpc$, the normalised version, are set out in equations 6.7 and 6.8 respectively. There is a special case for noun synsets at the very top levels of the hierarchy. Given their prominence in all texts, all nodes at depth 1 and 2 have been assigned a value of 0.001. A second special case is that of adjectives. The organisational structure of adjectives in WordNet (see section 3.2.3) is significantly different from that of either nouns or verbs which share certain similarities. As such, the same measures could not be applied to calculate a weighting for adjectival nodes. For this reason, adjectives are considered to fall into one of two classes: core adjectives, those which form the centre of clusters, or satellites. Core adjectives are assigned a value of 0.5 and satellites a value of 0.001. This effectively prioritises core adjectives and only those satellites

which are highly frequent will impact on the overall WNDag weight. As all satellites that appear in text introduce to the WNDag object for that text the core adjectives to which they are related, this measure clusters adjective weights on core adjectives. A possible further step would be to ignore satellites completely and transfer all features to the core adjectives which could reduce the dimensions of the graphs significantly. These values set for special cases were derived from an estimation of the most effective values in my experiments.

$$Spc = \frac{(depth + \ln(sisters) - ln(branchingFactor))}{19} \tag{6.7}$$

$$NSpc = \frac{\frac{depth}{MAX_{depth}} + \frac{\ln(sis)}{2 \times \ln(MAX_{sis})} + \frac{(1 - \frac{\ln(BF)}{\ln(MAX_{BF})})}{2}}{2} \tag{6.8}$$

The maximum values for depth, sisters and BF (branching factor) in version 2.0 of WordNet are 16, 798 and 393 respectively. As the range of the sis and BF values is quite wide and very much skewed towards lower values (see section 4.2.5 and 4.2.7), the log of both values and their respective maxima where applicable are used. The maximum *Spc* value is 0.945 and the minimum value is 0.[5] The *NSpc* value ranges between 0 and 1, where 0 is completely general and 1 very specific. Tables 6.4 and 6.5 set out several examples of *Spc* and *NSpc* values respectively for specific WordNet synsets. It would appear from the specificity coefficients for "woman" and for "horse" that the normalised specificity measure is more successful in satisfying pre-theoretic intuitions of concept specificity. For a given Arc, the node specificity value for that Arc is the sum of the specificity values of its source and

[5]Where the denominator for the *Spc* value is less than or equal to 1, the denominator is set to 1. This is to avoid calculating negative specificity coefficients and to control for errors from attempting to divide by 0.

target NodeInArc objects. The next section discusses how these features are combined to calculate weights for WNDag objects.

Synset	**Depth**	**Sis**	**BF**	$\ln(depth * sis)$	$\ln BF$	Spc
organism	4	4	41	2.77	3.71	0.07
person	5	55	358	5.6	5.88	0.09
woman	7	6	59	3.73	4.07	0.093
horse	13	6	28	4.35	3.33	0.13
potter	9	34	2	5.7	1	0.57
swan orchid	11	87	0	6.86	1	0.68

Table 6.4: Node Specificity Examples

Synset	**Depth**	**Sis**	**BF**	$\frac{depth}{MAX_{depth}}$	$\frac{\ln sis}{\ln MAX_{sis}}$	$(1 - \frac{\ln BF}{\ln MAX_{BF}})$	$NSpc$
organism	4	4	41	0.25	0.2	0.38	0.019
person	5	55	358	0.31	0.59	0.015	0.003
woman	7	6	59	0.4375	0.268	0.32	0.118
potter	9	34	2	0.5625	0.53	0.883	0.264
horse	13	6	28	0.8125	0.27	0.44	0.09
swan orchid	11	87	0	0.6875	0.669	1	0.459

Table 6.5: Normalised Node Specificity Examples

6.6.2 System techniques for weight calculations

The eight features set out in the previous section can be combined in many different ways. The following core research objectives have been addressed in this implementation and are the focus of the performance evaluation experiments:

- exploiting all of the information available in the text, in the form of repetition and paraphrase counts;
- mitigating the effects of polysemy (both in texts and in sister terms in synsets);
- prioritising different arc types;
- using a measure of node specificity to normalise weights relative to WordNet topology.

The system I have prototyped allows techniques employing different combinations of features to address these aims. Table 6.6 sets out the features used by each of the seventeen techniques.[6] The technique formulae are set out in table 6.7.

The techniques can be divided into three groups:

1. text features (1–9);
2. WordNet features (10–14);
3. text and WordNet features (15–17).

The first technique assigns a weight of 1 to all Arcs regardless of traversal frequency or any other feature. This amounts to a baseline count of absence/presence of an arc in a WNDag object. All other techniques include the Arc traversal count as primary feature. Techniques 2–5 combine traversal count with a single text-based frequency count feature. As stated above the effect of *inText* counts, when put to composite use, is multiplicative on arc

[6]The feature names in the table refer to the abbreviations listed in table 6.2. "X" in a cell indicates the features used in this technique combination. The empty cells provide much scope for future work on different feature combinations.

	Technique	Frq	Txt	Rep	Par	Pol	Rel	Spc	NSpc
1	Count								
2	Frequency	X							
3	InText	X	X						
4	Repetition	X		X					
5	Paraphrase	X			X				
6	NoPolysemy	X				X			
7	FactorOutPoly	X				X			
8	InText NoPoly	X	X			X			
9	AllText	X		X	X	X			
10	ArcType	X					X		
11	NodeSpec	X						X	
12	NormNodeSpec	X							X
13	ArcTypeSpec	X					X	X	
14	ArcTypeNormSpec	X					X		X
15	ArcTypeAllText	X		X	X	X	X		
16	All	X		X	X	X	X	X	
17	AllNorm	X		X	X	X	X		X

Table 6.6: System Techniques for Weight Calculation

	Technique	Formula
1	Count	$weight = 1$
2	Frequency	$weight = Frq$
3	InText	$weight = Frq + Txt$
4	Repetition	$weight = Frq * Rep$
5	Paraphrase	$weight = Frq * Par$
6	NoPolysemy	$weight = Frq/Pol$
7	FactorOutPoly	$weight = \frac{Frq-Pol}{NodePoly}$
8	InTextNoPoly	$weight = Frq + \frac{Txt}{Pol}$
9	AllText	$weight = Frq * \frac{(Rep*Par)}{Pol}$
10	ArcType	$weight = Frq * Arc$
11	NodeSpec	$weight = Frq * Spc$
12	NormNodeSpec	$weight = Frq * NSpc$
13	ArcTypeSpec	$weight = Frq * Arc * Spc$
14	ArcTypeNormSpec	$weight = Frq * Arc * NSpc$
15	ArcTypeAllText	$weight = Frq * Arc * AllText$
16	All	$weight = Frq * Arc * Spc * AllText$
17	AllNorm	$weight = Frq * Arc * NSpc * AllText$

Table 6.7: System Techniques Formulae

frequency. The purpose of this is to assign greater value to those elements of the WNDag object that are present in the source text while the weight for all other elements for these techniques is equal to the arc traversal count alone. These techniques are intended to test the effect of text choices, in particular, the effects of repetition and paraphrase as cohesive devices, on the overall weight of WNDag objects.

Polysemy count, although derived from WordNet, has been included as a text-based feature as it is a function of lexical choices rather than WordNet semantics or topology. Technique 6 is the first attempt to handle polysemy. It is a simple division of arc traversal count by the sum of the polysemy counts (if any) of its source and target NodeInGraph objects. Technique 7 reduces the arc frequency count as a function of the polysemy count for its NodeInGraph objects. Techniques 8 and 9 aim to constrain the treatment of polysemy specifically to the counts of those polysemous terms that appear in text, as described in section 6.6.1. Repetition and paraphrase counts for each term are divided by the polysemy count for that term. These techniques are intended to mitigate the effect of polysemous words on the overall WNDag object count.

Techniques 10–14 address the features derived from WordNet structure: relation type and the two types of node specificity. As noted in section 6.6.1, the Node specificity value is the sum of the specificity values for the source and target NodeInGraph objects. Both the relation type and node specificity values range from 0 to 1. Therefore, their effect is to reduce Arc weight. Techniques 10–12 test each feature in turn. Techniques 13 and 14 test relation type with the two types of node specificity value in turn. These techniques are intended to factor in the structure of WordNet to the calculation of WNDag object weight.

Technique 15 combines the relation type weighting algorithm with the AllText comprehensive text-based cohesive devices measure. Techniques 16 and 17 combine both node specificity and relation type features of WordNet structure with features of textual cohesion and mitigate the total weight with regard to polysemy in the source text. They differ only in the selection of node specificity value.

The seventeen techniques listed aim to test all aspects of the representations described in chapter 5. At its most basic level, an intended contribution of the work described here is to address the problem of data-sparseness for text-based (that is, explicit word– or lemma– based) measures of the informational contribution of texts. The use of additional level semantic tagging supplied by WordNet nodes provides the extra data necessary to measure semantic similarity. Techniques 1 and 2 test the effectiveness of a simple inclusion of knowledge base information. It is clear, however, that representations based on non-disambiguated texts suffer from excess noise. The remaining techniques, therefore, aim to reduce this noise with recourse to features derived from the text and derived from the knowledge base. The relative advantages or disadvantages of each techniques will be discussed in detail in chapter 8. The following section of the current chapter discusses how the weights calculated according to the different system techniques are evaluated to determine significant similarity of texts.

6.7 Determining significance

The previous section discussed how the overlap and the difference between two WNDag objects derived from text is measured. This section discusses how these measures are evaluated to determine whether the size of the overlap

is of significant size to warrant a claim that the two source texts are similar. Figure 6.1 shows a simple Venn diagram of two intersecting sets. The problem we are faced with here is to determine whether $A \bigcap B$ contributes more to the overall information content of $A \bigcup B$ than one would expect according to chance. The key factor here is to determine what is the distribution which might be expected according to chance. Measures that may be used are discussed at length below.

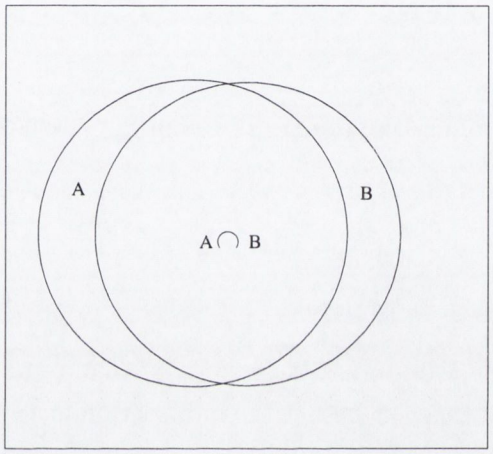


Figure 6.1: Venn Diagram with Large Set Intersection

6.7.1 Possibilities for set intersection

Prior to this, however, it is useful to look at the question of set intersection in abstract and to illustrate the different intersection possibilities which may arise.[7] Figure 6.1 presents a simple scenario where the sets are of more or less equal size and it would appear that the set intersection comprises a

[7]In this exposition, I will discuss intersection in terms of sets illustrated with Venn diagrams rather than as graph intersection.

significantly large portion of both sets. Figure 6.2 shows the opposite case where the intersection is minimal relative to the size of both sets. These two

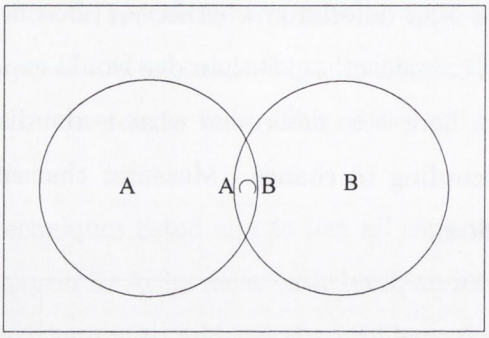


Figure 6.2: Venn Diagram with Small Set Intersection

cases may be carried to extremes where the sets either overlap completely ($A \bigcap B = A = B$) or do not overlap at all ($A \bigcap B = \emptyset$), as in figures 6.3 and 6.4. Figure 6.5, however, shows a more common situation where the intersection is large but it is not clear whether the intersection is a significantly large portion of either set. The similarity measures discussed below are designed to identify the size of intersection which is of significant size relative to the size of the two input sets to suggest that the two sets are similar.

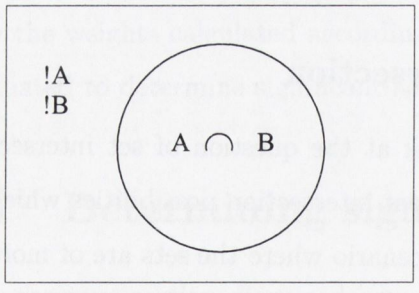


Figure 6.3: Venn diagram with no set difference

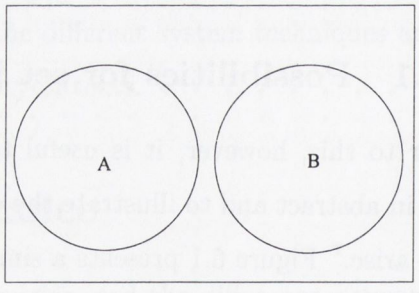


Figure 6.4: Venn Diagram with no set intersection

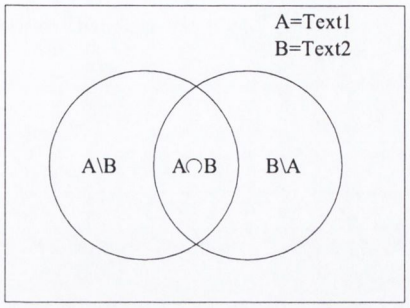


Figure 6.5: Venn Diagram with Moderate Set Intersection

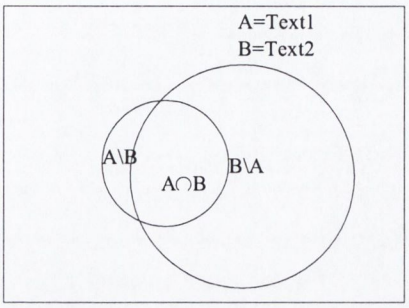


Figure 6.6: Venn Diagram with Set Subsumption

A more problematic case is illustrated by figure 6.6 where the cardinality of the sets is very different. Clearly, the set intersection is very significant relative to the size of set A but not relative to the size of set B. The measure of intersection significance must be sensitive to the possibility of subsumption illustrated by this example. The similarity measures that may be employed to determine significant overlap between Storyline text representation graphs are set out in section6.7.2.

6.7.2 Baseline similarity

A secondary point to note in this discussion is the possibility of baseline similarity between texts, particularly texts of the same genre. One might expect that the news texts in the RTÉ, Australian news and Google news corpora would have a baseline similarity as they are all modern news texts. This baseline similarity might be indicated by some shared WordNet synsets. Surprisingly, however, an experiment to measure baseline similarity in the 25 gold standard RTÉ corpus texts (see section 7.3.4) showed that, in fact, the

WNDag graph objects representing the input texts shared only the following four NodeInGraph objects derived from WordNet:

- Psychological features (top level)
- Cognition (level 2)
- Abstraction (top level)
- Entity (top level)

This somewhat unexpected result suggests several conclusions. First, the graphs derived from WordNet are not sensitive to genre specific features of text. This is not surprising as the aim of the representation is to abstract away from the text to a certain extent. The transition from word in text to WordNet synset factors out the nuances of vocabulary selection focusing on semantic content. Second, it is unadvisable to ignore any element in the graph intersection as none of the intersection is common to all texts, it would appear. The fact that it is not possible to factor out certain more generic concepts *a priori* highlights the importance of the weighting algorithm. The algorithm should strike a balance between reducing the prominence of synsets representing generic concepts while acknowledging their potential importance. The final tentative conclusion is that a similarity measure based on comparison of graphs derived from WordNet should be a fair indicator of semantic similarity as spurious similarity does not appear to be an artefact of the hierarchical structure of WordNet.

6.7.3 Similarity Measures

The task of determining similarity is common in Information Retrieval where usually it is a question of determining the similarity of two term vectors.[8] This section discusses some of the measures in common usage and their appropriateness to the Storylines system reported in this thesis. All the measures discussed below are intended to evaluate the distribution of elements in intersecting sets, such as those in figures 6.1 to 6.6. Somewhat surprisingly, however, they may give very different results when employed on the same data sets (Rorvig 1999). Therefore, it is important to establish which measure is most appropriate to the Storylines implementation of text comparison. At their most basic level, all measures are based on a four-way table, such as that set out in table 6.8, indicating presence or absence in sets A and B. The meaning of a, b, c and d is as follows:

- **a**: elements present in both A and B ($A \cap B$)
  the intersection WNDag object
- **b**: elements present in A but absent from B ($A \backslash B$)
  the A less B difference WNDag object
- **c**: elements present in B but absent from A ($B \backslash A$)
  the B less A difference WNDag object
- **d**: elements absent from both A and B ($\neg A \cap \neg B$)
  the rest of WordNet

Other similarity measures, such as the Russel and Rao coefficient, simple matching, Rogers and Tanimoto coefficient and Sokal and Sneath measures,

[8] Many discussions of similarity measures for information retrieval exist in the literature (Rijsbergen 1989, Salton and McGill 1983, Chowdhury 2004).

		Text1	
		Presence	**Absence**
Text2	**Presence**	a	b
	Absence	c	d

Table 6.8: Four-way table for similarity measures

were not implemented as the calculation includes a count of joint absences, **d**, which, although it is possible to determine, in this instance, always far outweigh the presences and, hence, skew results. The measures listed below deal only with those elements which are present in the set, i.e. **a**, **b** and **c**. They differ, firstly, in the weight they assign to the intersection and, secondly, in their underlying assumptions regarding symmetry of the size of the two data sets.

The Dice and Jaccard Coefficients

$$\frac{2a}{(a+b)+(a+c)} = \frac{2a}{2a+b+c} = \frac{a}{a+\frac{b}{2}+\frac{c}{2}} \tag{6.9}$$

$$\frac{a}{a+b+c} \tag{6.10}$$

The formula to calculate the Dice coefficient is set out in equation 6.9, that for the Jaccard coefficient, also called the similarity ratio, is set out in equation 6.10. The values for the two coefficients range between 0 and 1, from completely dissimilar to identical. Both coefficients exclude joint absences from consideration, as this application requires. The difference between the two lies in their treatment of the intersection of the two sets. In the formula for the Dice coefficient, matches, i.e., the intersection, are given double weight (2a), whereas for the Jaccard measure, matches and non-matches are

given equal weight.

The Cosine coefficient

$$\frac{\sum_{k=1}^{t}(element_{ik} \cdot element_{jk})}{\sqrt{\sum_{k=1}^{t}(element_{ik})^2 \cdot \sum_{k=1}^{t}(element_{jk})^2}} \tag{6.11}$$

The Cosine measure is better expressed in terms of vectors rather than sets. It is a measure of the angle between two vectors of equal dimensions. For two vectors of length t, $Text_i$ and $Text_j$ representing two texts, the following terms are required to calculate the cosine measure for the two texts

- $\sum_{k=1}^{t} elements_{ik}$: the sum of all elements in $Text_i$, which corresponds to $a + b$
- $\sum_{k=1}^{t} elements_{jk}$: the sum of all elements in $Text_j$, which corresponds to $a + c$
- $\sum_{k=1}^{t} elements_{ik} \cdot elements_{jk}$: the sum of the products of corresponding elements in each vector, which corresponds to a

The formula for the Cosine coefficient is set out in equation 6.11. Unlike the Dice and Jaccard measures, the denominator is not additive. In fact, elements which occur multiple times in either set have greater prominence as, for the two sets, the count of each element is squared. These values are summed for each vector and the results are multiplied. The final result is the square root of this value.

The Asymmetric coefficient

$$\frac{a}{a+b} \text{ or } \frac{a}{a+c} \tag{6.12}$$

The final measure which was considered is the Asymmetric coefficient, set out in equation 6.12. In this instance, the size of the intersection is compared to the size of only one of the sets. The measures mentioned above have the disadvantage that if the cardinality of the two sets being compared are very different, the coefficient cannot reflect that the contribution of the intersection may be greater for one set than for the other. Figure 6.6 illustrates such a case where the intersection is clearly of vital importance to set A but not set B. As the denominators of the previous measures **combine** the cardinality of the two sets, whether as an additive or multiplicative function, this distinction cannot be drawn. The asymmetric coefficient, on the other hand, allows for this possibility. Indeed, Tversky (1977, p.328) argues that, contrary to traditional assumptions, "similarity should not be treated as a symmetric relation", given empirical evidence of asymmetric similarities. It is appropriate, therefore, that a similarity detection algorithm allows for asymmetric similarity judgments.

6.7.4 Implemented Measures

In order to investigate the effects of selecting a particular measure, the Jaccard, Dice and Asymmetric measures were implemented in this system. The Cosine coefficient was not implemented as the computation involved is significantly different from the other three measures and, therefore, it was not strictly comparable. The Jaccard and Dice measures are evaluated as baseline similarity measures while the Asymmetric measure is intended for use when the cardinality of the sets is substantially different. The results for the three measures are discussed in detail in chapter 8.

6.8 Identifying differences

> No Difference Without Similarity (Gentner and Markman 1994, p.152)

The identification of text differences must follow positive determination of text similarity. According to Gentner and Markman (1994, p.152) "it should be easier to find the differences between pairs of similar items than to find the differences between pairs of dissimilar items." This may seem counter-intuitive. However, if it is a question of selecting salient differences between dissimilar texts, in a sense, any extract from the text is sufficient. Whereas if the texts are similar, differences stand out against the backdrop of similar information. For this reason, the task of difference identification is only carried out in the event that the test for similarity discussed above yield positive results.

The algorithm for difference detection is set out in algorithm 6.4. The task described here involves identifying the *sentences* which contribute most to the difference WNDag object relative to their contribution to the intersection WNDag object. The task could be formulated as one of identifying concepts or arcs which are most salient to the graph object, however, sentences provide the most easily interpretable output to the system.[9] The current task of differential sentence identification is carried out by traversing the difference WNDag object, DAG_i that represents the components of some text, $Text_i$, which are not shared with a second text, $Text_j$. A list of weights is compiled for each input sentence of $Text_i$, calculated according to the same technique p used to measure text similarity. $Weight_{sent}$ for some

[9]The future work section 9.2.5 outlines possible directions for interpretable and informative salient difference output.

Algorithm 6.4 Identifying Key Differences

Input: Three graphs $G_{common} = (V_{common}, E_{common})$, $G_{1less2} = (V_{1less2}, E_{1less2})$, $G_{2less1} = (V_{2less1}, E_{2less1})$ where V_x is the set of vertices and E_x is the set of edges of G_x

Input: weight calculation technique, **p**

Output: $diffSentences_{text1}$: the 2 most different sentences in text 1

Output: $diffSentences_{text2}$: the 2 most different sentences in text 2

1: instantiate arrays **1Less2Sentences, 2Less1Sentences** to length of $Text_1$ and $Text_2$ respectively
2: **for** $G_i = G_{1less2}, G_{2less1}$ **do**
3: **for all** e_i such that $e_i \in E_i$ **do**
4: calculate weight w_i of e_i according to parameter **p**
5: **for** $sent = 0$ to $numberOfSentencesInText_1$ **do**
6: $graphSentences[sent] += (w_i * sentencesIn[i])$
7: **end for**
8: **end for**
9: **end for**
10: **for all** e_i such that $e_i \in E_{common}$ **do**
11: calculate weight w_i of e_i according to parameter **p**
12: **for** $sent = 0$ to $numberOfSentencesInText_1$ **do**
13: $1Less2Sentences[sent] -= (w_i * jointSentencesIn[sent][0])$
14: **if** $1Less2Sentences[sent] > maxDiffVal$ **then**
15: Add $sent$ to $diffSentences_{text1}$, $maxDiffVal = 1Less2Sentences[sent]$
16: **end if**
17: **end for**
18: **for** $sent = 0$ to $numberOfSentencesInText_2$ **do**
19: $2Less1Sentences[sent] -= (w_i * jointSentencesIn[sent][1])$
20: **if** $2Less1Sentences[sent] > maxDiffVal$ **then**
21: Add $sent$ to $diffSentences_{text2}$, $maxDiffVal = 2Less1Sentences[sent]$
22: **end if**
23: **end for**
24: **end for**
25: Output $diffSentences_{text1}$ and $diffSentences_{text2}$

sentence, is the sum of weights of each arc, Arc_i contained in DAG_i, which derives from a given sentence in $Text_i$. This process favours longer sentences as they are likely to contribute more arcs and nodes to a text graph. In order to address this bias and to ensure that the sentence does indeed contribute significantly to the difference graph, it is necessary to compare its contribution to the difference graph with its contribution to the intersection graph. The sentence weight list is, therefore, compiled for the intersection graph. The *differential* contribution for a sentence i is calculated according to the formula in equation 6.13.

$$differenceWeight_i - intersectionWeight_i \tag{6.13}$$

This simple formula gives a measure of the actual contribution of each sentence to the difference graph. The system then outputs the sentences with the highest differential contribution value. A comprehensive evaluation of both components of the Storylines text differentiation system is discussed in Chapter 8.

6.9 Discussion

This chapter describes the text differentiation component of the Storylines system I designed and implemented. The task of text differentiation in this approach comprises two sub-tasks: similarity estimation and differential content identification. The development of this system has addressed several of the research aims of this thesis, notably the full exploitation of the Storylines text representation for text differentiation. This aim has been achieved in the following ways:

- Section 6.6.1:

- elaborates how the contribution of the lexical cohesive devices of repetition and paraphrase can be quantified for individual nodes in the text model;
- outlines a weighting algorithm for WordNet relations based on repeated experimentation on the similarity estimation task;
- describes how the empirical results of the concept specificity experiment, set out in section 4.3, have formed the basis of two formulae for node specificity metrics in WordNet;

- Section 6.6.2 defines how these relational and topological features of WordNet as well as the measures of text-based cohesive devices in the Storylines text model can be exploited alone or in combination to quantify the informational contribution of individual model components. To this end, seventeen techniques combining different features of the model are enumerated. Chapter 8 sets out a thorough evaluation of these techniques on the tasks of similarity detection and differential content identification.

An ancillary contribution of the research reported in this chapter is an exposition of similarity measures in the literature and the implementation of three, the Jaccard, Dice and Asymmetric measures, for the similarity estimation task. These three measures are evaluated as part of the Storylines evaluation process reported in chapter 8.

Chapter 7

Gold Standard Human Evaluation

7.1 Introduction

Despite the fact that similarity is a key concept underlying many information retrieval and extraction systems, few empirical studies, such as that of Lee *et al.* (2004) and Ellman (2000), have been conducted to elicit similarity judgments from human evaluators. As computational measures of semantic similarity between texts is a central issue in this thesis, it was necessary to examine human judgments and if possible to create a gold standard of human judgments against which to evaluate the performance of the system described in chapter 6.

In this chapter, I describe the psycholinguistic study carried out to elicit judgments from human respondents as to the similarity of pairs of texts drawn from the RTÉ news corpus described in section 3.3. Section 7.2 sets out the motivations for the study: the research goal and expected application of the study results. Section 7.3 describes aspects of the methodology used to undertake this study:

- details of the system implemented to administer the study (7.3.2)
- participants (7.3.3)
- materials: the texts used and their selection process (7.3.4)
- procedure (7.3.5)

Section 7.4 outlines some basic results of the study regarding numbers of participants, response times and distribution and range of ratings. Section 7.5 provides an in-depth analysis of methods for assessing reliability of experiment results and their application to this experiment.

7.2 Motivations

As noted above, the question of semantic similarity is central not only to this thesis but to many information processing applications such as information retrieval and text categorization. While hierarchical organisation of databases or file systems or chronological ordering of news stories may give an indication of similarity among text documents, they can only provide indirect access to human similarity judgments. Therefore, although the cost, in terms of effort, of gathering empirical data of pair-wise semantic similarity judgments becomes prohibitive for a large corpus of texts, it is imperative that a subset of texts be used to carry out such a study in order to provide validation criteria for any computational model of similarity.

The primary test corpus in this research is the RTÉ Interactive corpus. As noted in section 3.3.1, this corpus has some internal structure whereby texts relating to particular news events are grouped together by naming conventions. For example, two stories about Iraq may be named *Iraq1* and *Iraq2*. However, using file names as an indication of semantic similarity has several drawbacks.

1. The naming of files was carried out by one individual and hence is completely **subjective**.

2. The file names are **arbitrary**, although the assumption is that they convey something of the content of the text.

3. The most important drawback is that the relationship between files of the same name is **undefined**

    - Are they similar?
    - Are they chronologically related?

- Do they relate different facets of the same event?

Furthermore, even if naming conventions do delineate groups of texts which are semantically similar, there is no relationship defined between the different groups: they may all be equally dissimilar or some may be more similar than others. The study described in this chapter was carried out to establish a measure of similarity between a random selection of the texts in the RTÉ corpus. This similarity measure is intended to establish a gold standard against which to evaluate the model of similarity which emerges from the system described in this thesis.

7.3 Method

7.3.1 Experiment design and the talk-aloud protocol

During the design phase of the experiment, several questions arose as to an appropriate presentation format and in what manner to elicit the similarity ratings. In order to clarify these questions, a pilot version of the experiment was conducted as a talk-aloud protocol. The purpose of a talk-aloud protocol is to pilot an experiment design with a group of volunteers who together talk through the process of completing the experiment. This process highlights potential complexities of the experiment design. It also serves to ascertain whether the experiment is in fact eliciting expected subject responses and/or behaviour. The current talk-aloud protocol was conducted under supervised conditions with two volunteers participating together. The volunteers were given the task of comparing pairs of texts and giving a similarity rating. The complete materials are presented in appendix D.1. In brief, the similarity rating task was presented in the three formats in 7.1, 7.2 and 7.3 below.

(7.1) Volunteers were asked to give a similarity rating on the following horizontal numeric scale;

1 2 3 4 5 6 7 8 9 10

Identical .. .. Totally different

(7.2) Volunteers were asked to give a rating by selecting from the following vertical list of graded terms;

- **Identical**
- **Very similar**
- **Quite similar**
- **Quite different**
- **Very different**
- **Completely different**

(7.3) Volunteers were given the following list of binary questions and asked to select an answer for each.

- Are these two stories the same?
- Are these two stories about the same event?
- Are these two stories about the same type of event?
- Are these two stories on the same topic?
- Are these two stories the same genre?
- Are these two stories similar?

The volunteers were encouraged to discuss the process of deciding on a rating, what they found difficult and how the task could be rendered more simple and clear. They were recorded on minidisc by the researcher as they carried out the task. In further discussion with the volunteers, the design of the experiment was refined with regard to training, the rating scale employed and the distinction between similarity and relatedness.

A training page: to have or to have not?

News texts like all text, although all of the same genre, may be deemed to be similar in a number of different ways. Two texts may share similar news events, topics, themes, domains, styles and so on. There is a continuum of possibilities from fully identical texts to texts that differ completely in all aspects of content and form, such as a Kavanagh poem and a tax return form. It emerged from the talk-aloud protocol that these multiple layers of similarity were a source of confusion with respect to the basic task of giving a single similarity rating. When asked to judge similarity the volunteers asked themselves: "similarity of what aspect of the text?" For this reason, a training page was included in the final experiment.[1] The initial training phase is intended to encourage respondents to consider different aspects of text similarity and then give a rating. This process acknowledges and highlights that texts can be similar in many ways and that an overall similarity judgment is a related but somewhat separate issue.

The Rating Scale

The volunteers found that the list of binary questions (format 7.3 above) was helpful for coming to terms with the task of giving a similarity rating.

[1] A screenshot of the experiment training page is included in appendix D.2.

However, they both found that a ranked judgment, as in 7.1 and 7.2 above, was ultimately quicker and easier to provide. One volunteer in particular criticised the numeric scale (format 7.1) as she said people tended to avoid the extremes of the scale and focus on one or two values in the middle. Format 7.2, the list of terms in rank order from identical to completely different, proved to be the most popular. The volunteers did not have difficulty interpreting the terms. They also stated that they felt more definite about their ratings on this scale than on a seemingly arbitrary numeric scale. One volunteer also commented that the list of seven terms was too restrictive and should be supplemented with half-measures.

It should be noted that magnitude estimation was considered as a means of eliciting similarity judgments. Magnitude estimation entails assigning a number to each new stimulus, in this case, a pair of texts, relative to a standard stimulus, an initial pair of texts. The advantage of this technique is that, in theory, respondents assign ratings as a ratio to the initial stimulus. Therefore, all ratings can be scaled and comparison of ratings is not subject to questions of subjectivity regarding the scale used. Although this method has been used with some success in psycholinguistic experimentation, it does require that respondents are comfortable performing mental arithmetic in order to complete the questionnaire. Furthermore, during an informal talk-aloud protocol for a grammaticality judgment experiment using magnitude estimation in which I participated, subjects commented that they did not assign values in the magnitude estimation task based on ratios but, rather, used it as a boundless scale. It is not clear, therefore, whether the basic premise of the scalability of the technique is well-founded. For these reasons, I decided not to adopt this approach for this experiment.

Similarity *versus* relatedness

A final issue that emerged during the talk-aloud protocol was the interaction of the concepts of text relatedness and text similarity. This question arose in the context where two texts were deemed to be not very similar but the events recounted in the texts were strongly related. Text pair comparison D.13 in appendix D.1 prompted this discussion. The first text reports on proposals announced by Ariel Sharon for the Palestinian state. The second text reports on a series of gun battles in the Gaza Strip. The texts are highly related, however, they are not similar as one is an account of political manoeuvres in a political journalism style and the other an account of violence and military incursions in a more sensationalist style. This instance caused the volunteers some difficulty in that they felt that a low similarity rating was inaccurate because the texts were highly related. Conversely, a high similarity rating was inaccurate because the texts were not in fact similar. It became clear that in order to elicit a *similarity* rating in these cases, it was necessary to separate out the notions of similarity and relatedness. For this reason, the following parallel *relatedness* scale to the similarity scale was included in the final experiment:

(7.4)
- **Highly related**
- **Very related**
- **Quite related**
- **Quite unrelated**
- **Very unrelated**
- **Completely unrelated**

It was not mandatory to assign a relatedness rating. However, the relatedness scale provided a means of resolving the kind of problematic interaction

between the notion of similarity and relatedness that occurred during the talk-aloud protocol. In fact, the relatedness judgments elicited in this experiment provided a key insight into the subtleties of semantic similarity in text.[2]

7.3.2 Implementation

I implemented a CGI program to conduct the experiment remotely over the Internet. Like the Trinity on-line experimenter (Kenny 1998) or Web-Exp (Keller *et al.* 1998), it allows the researcher to conduct an experiment over the web and record its results. This program is customized to the requirements of this research.

Internet experiments

Web-based experiments have several well-documented pros and cons (Hewson *et al.* 2003). This study required over 200 respondents and had no monetary resources. I therefore decided to avail of the supposed higher response rate of subjects for web-based experiments and accept the associated disadvantages of limiting the potential subjects to Web users and the difficulty of subject authentification (see 7.3.3). A positive consequence of the web implementation is the level of control it allows in the presentation of materials to respondents.

- The software ensures that six pages containing random pairs of texts are presented to each subject in a pre-defined order. The subject is required to assign a similarity rating and a relatedness rating for each

[2]See section 8.5.5 for a discussion of text similarity, topic similarity and relatedness for experiment design and natural language processing applications.

text pair. He/She may return to a previous page and change their responses if they wish. However, all responses and the time at which they were submitted are recorded so both original and revised responses are available for analysis. The original similarity and relatedness ratings were used in all subsequent analyses and compilations of the experiment data. The revision facility was implemented primarily therefore to give respondents the illusion of control while participating in the experiment.

- The software ensures that the subject responds to all required experiment items. The subject may not complete all pages in the experiment but any page they do submit must have all required fields completed. This facility was implemented in the aim of enforcing full completion of the rating tasks. As there was no monetary or any other incentive for respondents to undertake or complete the experiment, all efforts had to be made to ensure that those that did participate participated fully.

Recorded Results

The results for each respondent are written to separate files. The software assigns a unique user ID and records this ID and the respondent's IP address. On completion of each page, the local network timestamp, text pair and similarity and relatedness ratings are recorded.

7.3.3 Participants

In total, 321 people participated in the on-line experiment. Of these, 236 completed all six text pair comparisons of the experiment. In order to maximize participation among those petitioned, participants were not required

to give any personal details, for example an e-mail address. For that reason, the experiment data does not include any information as to the sex, age or handedness of participants. A larger study might aim to gather this information, however, as noted above, these details were not requested in order to maximize participation in the experiment. Only the IP address for each participant was recorded. However, an IP address is not necessarily a unique participant identifier as the IP address may be that of an intervening proxy host or an arbitrary number assigned to a dial-up user. Therefore, there is no way to determine from the address:

1. the user's identity
2. whether the user has participated in the experiment previously

This difficulty in verifying a user's identity is a considerable disadvantage of the looser experimental conditions of web experiments. However, as this experiment did not aim to impose restrictions on the selection of participants, as to age or sex for example, this issue was accepted as a minor shortcoming that should not effect experiment results.

7.3.4 Materials

The texts used were selected from the RTÉ Interactive corpus described in section 3.3. This corpus consists of 628 individual texts gathered over five days on-site in the internet division of the Irish National broadcaster, RTÉ at their head quarters in Dublin: 22 and 27 November and 4, 5 and 6 December 2002. The average length of the texts is 787 characters with a standard deviation of 645. As the range of text sizes is so broad, the twenty-five experiment texts were selected at random from 129 possible texts in the range 587 – 987 characters. The texts range in length from 110 to 160 words

and cover a broad range of topics in Irish and world news in late 2002. The texts are set out in appendix A.2. Five typing errors in the original twenty-five texts were corrected and a repeated headline was removed from two of the twenty-five. The distribution of the texts[3] according to the day they were collected is set out in table 7.3.4.

Date	# Texts
22/11/2002	4
27/11/2002	5
4/12/2002	10
5/12/2002	4
6/12/2002	2

Pair-wise combinations of the twenty-five individual news texts yields three hundred text pairs for comparison. One text pair was selected to be used in the training page also. Prior to conducting the experiment, the list of 300 text pairs was randomised. Ten independent ratings were elicited for each text pair. Each respondent was required to complete the training page and six text pair comparisons. The 300 pairs, therefore, were divided into sub-groups of six text pairs whereby each subgroup of six random text pairs was evaluated by the same ten respondents. This procedure was aimed to facilitate inter-rater agreement evaluation where agreement could be determined on a group by group basis. All experiment instructions and materials have been included as appendix D.

[3]A χ^2 test at 4 degrees of freedom shows the distribution according to day of collection is not significant.

7.3.5 Procedure

Participants were solicited by e-mail from among friends, family and colleagues of the researcher. Eighty people approximately were petitioned directly and some of these forwarded on the request to their friends and colleagues. Each participant accessed the experiment on-line with their web browser at:

www.cs.tcd.ie/Ann.Devitt/cgi-bin/ExperimentIntro.html

The CGI perl source code to administer the experiment on-line has been included in appendix D.3.

Instructions Page

On accessing the site, each participant was presented with an opening page which set out the instructions for the experiment.[4] The instructions state the provenance of the stories and, for participants unfamiliar with Ireland, the terms Gardaí, the Irish police force and Garda, a policeman, are explained. The instructions regarding the training page and subsequent pages are as follows:

> The training page: 1 pair of stories
>
> The first page is a training page to give you a feel for the task in hand. You are presented with 2 texts and asked five yes/no questions about what aspects of the stories are similar and five yes/no questions about what aspects of the stories are related.
>
> You are then asked to give a rating of how similar you think the stories are and a rating of how related you think the stories are.

[4]The screenshot of the instructions page is printed at Appendix D.1

> The choice of ratings is presented as a list starting at "Identical" or "most related" and going down to "completely different" or "completely unrelated". The checkboxes are in rank order so the blank in-between checkboxes may be selected even though they have no label associated with them.
>
> The experiment: 6 pairs of stories
>
> After this training page, there are 6 pairs of stories to compare. In these pages, you are only asked to give a similarity rating and a relatedness rating. You are not asked any yes/no questions.

Finally, there is an example given to illustrate the difference intended by the experimenter between the terms similar and related and the participant is requested to make this distinction where necessary.

Training phase

When the participants have read the instructions, they click on the start link taking them to the training page.[5] Participants are presented with two texts, one about a murder investigation in Kilkenny and the other about an attempted firebomb attack in Belfast. The texts are presented one above the other, the murder story on top and the firebomb story below. Participants are asked to read the two texts and answer the questions that follow. There follows a list of five yes/no questions on different aspects of similarity and relatedness of the two stories. The yes and no options are presented as radio buttons which the user may click yes or no to make their selection. In the default setting for these buttons neither option is selected and the user may select only one option.

[5]The screenshot of the training page is included as Appendix D.2.

Participants are then asked to give a similarity and a relatedness rating for the two texts. These tasks are proposed with questions 7.5 and 7.6.

(7.5) How similar are these two news texts?

(7.6) How related are these two news texts?

Each question is followed by a list of eleven radio buttons representing eleven possible ratings of which the participant may select only one. The list of possible ratings given is set out in table 7.1.[6] The instruction page notes

Similarity rating term	**Associated value**	**Relatedness rating term**
Identical	0	Highly related
..	1	..
Very similar	2	Very related
..	3	..
Quite similar	4	Quite related
..	5	..
Quite different	6	Quite unrelated
..	7	..
Very different	8	Very unrelated
..	9	..
Completely different	10	Completely unrelated

Table 7.1: List of similarity ratings with associated value

that the list is in rank order and the empty radio boxes may be checked. These empty checkboxes are provided to allow the participants more nuanced

[6]The associated numerical value for each rating given in column two of the table was not included in the actual experiment.

ratings as requested by the participants in the talk-aloud protocol. The only field that participants are required to complete is the similarity rating. They may not pass on to the next page until a rating has been assigned.

The Experiment

After this training phase, the participant enters the experiment proper. The next six pages each present a pair of news texts. The participant is asked to read the texts and to give a similarity rating and a relatedness rating for each pair. The rank ordered list of ratings are the same as in the training phase. Again, the only required field is the similarity rating and the user may not proceed to the next comparison until a rating has been assigned.

Participants are grouped in blocks of ten so that each new participant is assigned to a particular respondent group and the same six text pairs are presented to each member of this group. As there can be effects of ordering on responses, the six text pairs are presented in a fixed order so that any effects are consistent across respondents. Each pair is presented one above the other. The top-bottom ordering is randomly assigned at run-time. Each page states how many comparisons have been completed out of the required six. This facility was implemented in response to a suggestion from one of the talk-aloud protocol volunteers. She noted that the task was not an easy one, therefore, an indication of how close respondents were to finishing might spur them on to finish all six text pairs.

At any stage during the experiment, a participant may click the *Back* button on their browser to return to a previous page. They may choose to change their original rating and submit a new rating. The CGI script records all ratings that are made and the time they were made. Hence the original and subsequent revised judgements are available for contrastive analysis if

required. Details of subject revisions are set out in section 7.4.1. All other analyses are based solely on the original data.

7.4 Results

7.4.1 Some basic results

Participation

321 individuals participated in the experiment. Fifty-seven of these, 18% of respondents, only completed the training page. Of the remaining 82%, two completed four out of six comparisons, eleven completed two out of six comparisons and a further eleven completed only one of the six comparisons. 236 respondents, 74.4%, completed the entire experiment. The rate of attrition could be attributed to the unsupervised conditions of on-line experiments. Respondents are completely free to give up on the task if they are finding it difficult. They also are likely to be taking time from other activities or work to complete the experiment and therefore the probability of their being distracted from its completion is quite high.

Time

The completion time for each similarity rating task varied greatly among participants, from 4 to 1000 seconds. Some respondents commented in personal feedback to the experimenter that they had had to interrupt the experiment for some reason and return to it later. One can surmise then that the response times over about 300 seconds were due to people interrupting the experiment, rather than taking over five minutes to deliberate on a similarity rating. The histogram in figure 7.1 shows that the vast majority of ratings

were completed in under 250 seconds with an average response time of 75.7 seconds with a standard deviation of 174.5.

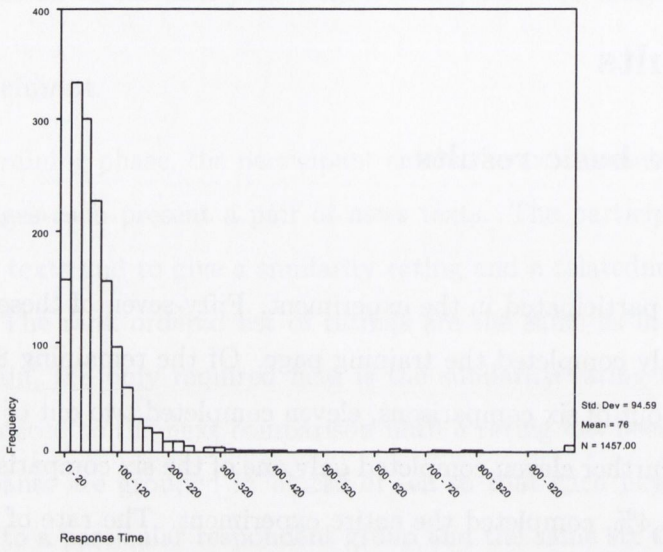


Figure 7.1: Histogram of response times, 20 second intervals

Revisions

Eleven people used the Back facility to revisit a text comparison they had already submitted. Of these, only three actually changed the similarity rating. One respondent changed the first rating three times from "completely different", rank 10, to "quite similar" rank 4, to rank 5. The other instances of changed ratings however were only single changes of one rank point. Similarly, only two individuals changed the ranking for the relatedness rating: one from 7 to 8 and the other from 2 to 4. The original ratings are used in all subsequent analyses.

Some feedback

The lack of changed ratings and the quite low average time taken to complete a text comparison task might indicate that participants found the task quite straight-forward. However, many participants commented that they found the task quite hard. The most common feedback comment was that the participant thought he/she was "wrong" or had approached the task incorrectly because he/she found all texts to be very, if not completely, different.[7] This feedback suggests three related observations:

- Inadequate instructions.

  The instructions given to participants may have been inadequate. Indeed, the instructions pages failed to stress that any rating given was purely a matter of opinion and that there was no right or wrong answer.

- Difficulty of the task.

  This negative feedback could also stem from the fact that respondents found rating texts for similarity a difficult task. As I noted in section 7.3.1, the multilevel nature of text makes the task of assigning similarity ratings a difficult one. Although the training page is intended to raise awareness of what similarity in texts entails, this may not have helped respondents in assigning ratings. Lee *et al.* (2004) report no such negative feedback on their similarity experiment where respondents are asked to judge similarity of *topic*. This raises questions as to how text similarity should be defined in such an experiment and whether a general notion of text similarity is too fuzzy for respondents to handle. Section 8.5.5 outlines some implications of the definition of similarity for experiment design and computational applications.

[7] Ellman (2000, p.119) reports the same feedback on task difficulty.

- The paradigm of the "good respondent"

  However, it is unclear whether this anxiety to give the "right" answer can be addressed by merely adapting experiment instructions. In the case of grammaticality judgment experiments, it was found that different instructions did not elicit a statistically significant difference in responses (Cowart 1997). It is possible that respondents strive to be "good" respondents regardless of the instructions given and it is difficult to affect a change in this behaviour. Schumann and Presser (1981) discuss the "acquiescence quagmire" in conducting experiments where some respondents, particularly from specific social groups, tend to agree with the individual conducting the experiment.

  Many of the respondents in this experiment had some connection with the researcher and this may have promoted their "good respondent" impulse and, hence, the fear of giving an "incorrect" response. However, this "quagmire" appears to be common in psychological experimentation in general. Therefore, the attendant respondent anxiety is, in a sense, irrelevant if the responses themselves are consistent across individuals.

7.4.2 Distribution of similarity ratings

Figure 7.2 shows how the distribution of ratings over all trials were heavily skewed towards low similarity ratings with a skewness value of -0.952 and a kurtosis coefficient of -0.082. The mean for all trials was 7.994. The "completely different" rating, 10, accounted for 45% of all ratings. The mid-points allowed on the scale were rarely used. In fact, only the exact mid-point of the range was used with any frequency, 2.7% of ratings. The original 11 point scale was therefore reduced to the 6-point scale set out in table 7.2.

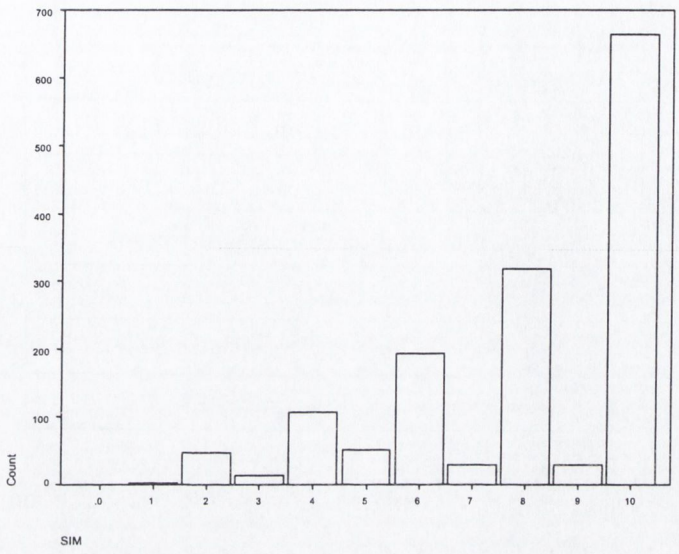


Figure 7.2: Bar chart of similarity ratings on full 11-point scale

Figure 7.3 shows the histogram of the ratings on this 6-point scale with the normal curve overlaid. It is clear from this figure that the choice of ratings do not follow a normal distribution.

7.4.3 Distribution of relatedness ratings

Overall, the ratings for degree of relatedness among the story pairs were very strongly correlated with the similarity ratings given. Figure 7.4 shows the distribution of values which follows that shown in figure 7.2 Again, the ratings were skewed towards low relatedness ratings with a greater number of "completely unrelated" than "completely dissimilar" ratings. The mean relatedness rating was within 0.01 of the mean similarity rating. Although the relatedness ratings did display slightly more variability with a standard

New Rating	Original rating and value
0	Identical(0) and mid-point (1)
1	Very similar (2) and mid-point (3)
2	Quite similar (4) and mid-point (5)
3	Quite different (6) and mid-point (7)
4	Very different (8) and mid-point (9)
5	Completely different (10)

Table 7.2: Reduced similarity rating scale with original value mappings

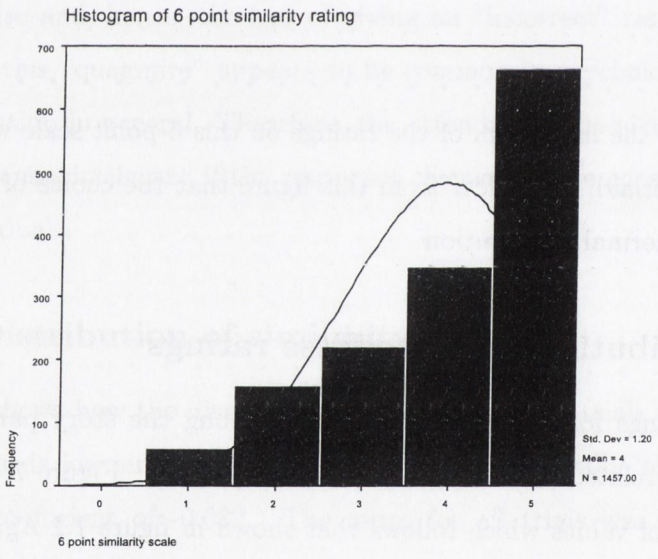


Figure 7.3: Histogram of similarity ratings on 6-point scale

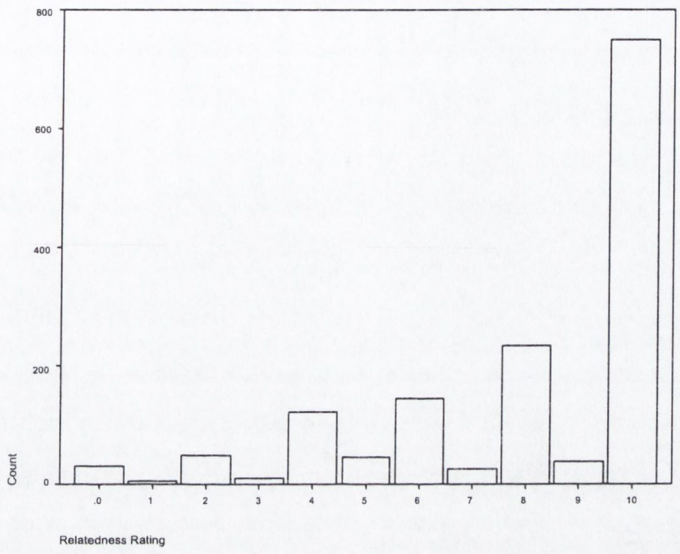


Figure 7.4: Bar chart of relatedness ratings on full 11 point scale

deviation of 2.7, compared with 2.3 for similarity ratings.

Section 7.3.1 outlines the reasons for including a relatedness rating scale in the experiment, notably that for certain text pairs there is significant interference between the two concepts. This interference could colour the similarity ratings if respondents had no means of separating out the notions of similarity and relatedness. In order to investigate whether this distinction is, in fact, motivated, an analysis of the difference between the two rating sets was carried out. On a story pair by story pair basis, the relatedness and similarity ratings are often very close, if not the same. The histogram in figure 7.5 shows the normal distribution of differences between the two ratings for all story pairs. Although the vast majority of ratings are within ± 2 of each other, some story pairs show a large difference between the sim-

ilarity and relatedness rating. The most notable difference between the two results is that at the "highly related" end of the relatedness scale there were 32 ratings, compared with 1 at the identical end of the similarity scale. The greater number of top ratings may be related to the fact that the top of the relatedness scale ("highly related") is not an absolute like the top of the similarity scale ("identical"). More interestingly, however, these highly related ratings did not necessarily correlate with stories deemed very similar. This discrepancy would suggest that similarity does not necessarily entail relatedness for respondents for some text pairs, as was suggested by volunteers in the talk-aloud protocol. Section 8.2.4 discusses some consequences of highlighting this distinction in testing the system described in chapters 5 and 6 and section 8.5.5 outlines some implications for future experiment design.

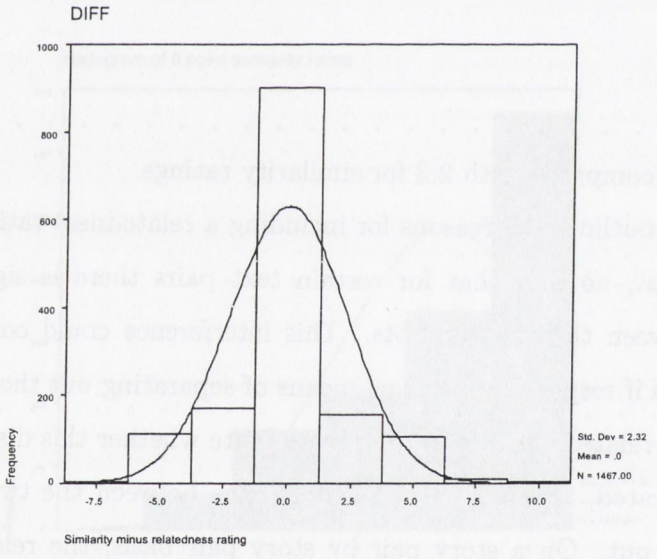


Figure 7.5: Histogram of difference between similarity and relatedness ratings

7.4.4 Range of ratings used by individuals

Respondents were presented an 11-point scale from which to select their response ratings for each of the six story pairs in the experiment. In fact, approximately 10% of respondents who completed the whole experiment used only one value for each response given. The average range of values was 3.68. A single individual made use of the full range of values (on the 6-point scale). A further 10% used five out of six ratings on the 6-point scale. Although it would appear from anecdotal feedback that respondents felt all story pairs were equally dissimilar, in practice this proved not to be the case as most respondents gave a spectrum, although perhaps not a broad spectrum, of ratings for the six story pairs they rated.

7.5 Reliability: Inter-respondent agreement

The purpose of this experiment was to build a gold standard of similarity ratings given by humans against which to test the system described in chapters 5 and 6. The system automatically estimates similarity of texts, even where the vocabulary is different, thus providing a tool for judging similarity and difference in context. The validity of a gold standard, however, depends on establishing some degree of consensus among the human respondents in the experiment as to the similarity of story pairs with which they were presented. The aim of this section is to determine the degree of agreement among respondents in the experiment described above.[8]

The task of determining agreement among ratings assigned by two or more raters is a common problem in many fields. The raters may be doctors

[8]In this section, I will use the generic terms "rater" and "rating" to denote the respondents and their similarity ratings.

and the ratings, their diagnoses on patients, or they may be marine biologists rating coloration of male stickleback fish, or students rating sentence grammaticality. The range of possible raters and ratings is as broad as the range of experiments or surveys that may be devised across the sciences. The basic issue across all disciplines, however, is the same: to determine whether the raters, whoever they be, agree on their ratings, whatever they be. The purpose of the original rating task will dictate the degree of agreement required. For example, diagnoses of mental disorders given by psychiatrists are often critical to a patient's future, therefore, circumstances would require that the experts are in near perfect agreement. On the other hand, a marketing firm conducting an experiment on the degree of positive response to a sensory stimulus, such as a colour, may not require perfect agreement among raters but the aim may rather be to determine some central tendency revealed by agreement patterns in the data.

This study falls into the same category as many studies in the field of linguistics where the research is intended to model or at least emulate human performance on a given task. The assumption is that human performance will provide the gold standard, the correct responses, against which to evaluate any system performing this same task. This assumption gives rise to two questions:

1. is this a task which humans can perform consistently?

2. is there an upper bound to consistency of performance among human raters?

In the case of word sense disambiguation, for example, Jorgensen (1990) found that percent agreement among human raters on a word sense disambiguation task was only 68%. In the light of these results, Gale *et al.* (1992)

propose an easier task of sense discrimination to Jorgensen's sense grouping experiment. They claim that if the task for humans is too hard, there will be much variability in their responses and, therefore, the responses are less useful for evaluating word sense disambiguation engines. In estimating the upper bound of this task, they found it necessary to redefine the task for human judges in order to attain higher agreement.[9]

This example illustrates that in order to model human performance on a task, it is necessary to determine that the task in hand is one which humans themselves perform reliably. Determining inter-rater agreement on rating tasks is intended to test this reliability by ensuring that raters' judgments are "reliable". Reliable, here, has the formal sense set out in Kaplan and Goldsen:

> The importance of reliability rests on the assurance it provides that data are obtained independent of the measuring event, instrument or person. Reliable data, be definition, are data that remain constant throughout variations in the measuring process. (Kaplan and Goldsen 1965, pp. 83-84)

Reliable data of judgments from raters rating the same information should be consistent and, hence, replicable.

Krippendorff (1980, p.120) states that reliability tests that the data reflect some reality, in this case, real human judgments of similarity of texts. The related test of validity, comparing data to some known standard or truth, is not required in this case as the intention is to derive this standard from reliable data. If the results of an empirical study, such as that described here,

[9]This association of task difficulty with low agreement among humans is reflected in the results of this study contrasted with that of Lee *et al.* (2004).

are reliable, they may be deemed a gold standard of performance on that particular task. Krippendorff further distinguishes three kinds of reliability:

1. Stability
2. Reproducibility
3. Accuracy

Stability relates to invariance of data over time, reproducibility relates to invariance of data under different test conditions and accuracy to data validity (i.e. how well the data conforms to expected or standard data). The form of reliability in question here is reproducibility. In this case, the different test conditions are the different raters undertaking the same task.

> Reproducibility is also termed "inter-coder reliability", "intersubjective agreement" or the "consensus" achieved among observers. (Krippendorff 1980, p.131)

These raters must be independent in order to give fully interpretable results. This experiment fulfils that requirement as each rater completed the task on their own computer unaware of other participants undertaking the same experiment.

The purpose of this section, then, is to determine inter-coder reliability by calculating a measure of agreement on ratings given for each story pair. Bishop *et al.* (1975) in their exposition of measures of association and agreement note that:

> when we look for a measure of association, there is rarely a need to develop a new measure. Rather the problem is to choose wisely from among the variety of existing measures. (Bishop *et al.* 1975, p.373)

The following sections set out five of the many different approaches to measuring inter-rater agreement available and discuss the results of the analysis on the experiment data. The approaches employed are as follows:[10]

- Deviation from mean and mode, measures of central tendency (section 7.5.2)
- Percent agreement (Gale *et al.* 1992) (section 7.5.3)
- the Kappa coefficient (Cohen 1960, Siegel and Castellan 1988, Krippendorff 1980) (section 7.5.4)
- the Cochran Q statistic (Cochran 1950) (section 7.5.6)
- Euclidean distance (section 7.5.7)

Section 7.5.5 sets out a justification for using the Cochran Q statistic on this data. Finally, section 7.5.8 discusses agreement in the training data.

7.5.1 The Data

As noted in section 7.3.2, each respondent's similarity and relatedness ratings were written to an individual file. These files were processed and the resulting data formatted as matrices of story pair by respondent ratings. The similarity and relatedness rating matrices for each group of respondents are set out in appendix E.1. Each group consisted of ten respondents who rated the

[10]Two related approaches which are notably absent are Latent Trait Analysis and Latent Class Modelling (Dillon and Mulani 1984, Uebersax 1999). In the absence of a gold standard, these approaches aim to extrapolate a gold standard from the data using factor analysis and subsequently test rater agreement with this factor. In the light of observations made and conclusions drawn as a result of applying the other approaches to the data, these analyses were not used on the data.

same six story pairs, although some members of the group may not have completed the ratings for all story pairs. There are 32 groups in total. The complete data is presented in appendix E.2 with two graphs for each group. Figures 7.6 and 7.7 present these graphs for group one. The first graph plots each rater's ratings for each story pair. The second plots the average rating assigned for each story pair in the group. These graphs give some indication of some general tendencies in the data. They illustrate the bias towards dissimilarity ratings and within this bias the degree of variability between users. Section 7.5.7 discusses intra-group variability and how it may be assessed. Subsequent sections deal with analysing the raw individual ratings data to gauge overall inter-rater agreement for the experiment.

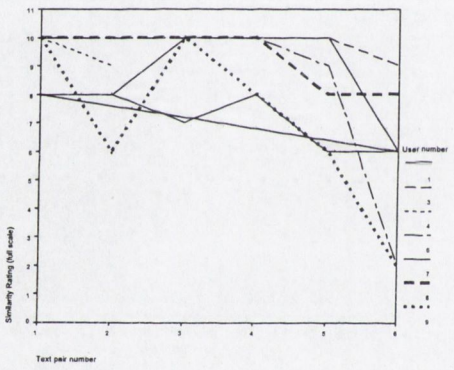


Figure 7.6: Similarity Ratings for Group 1

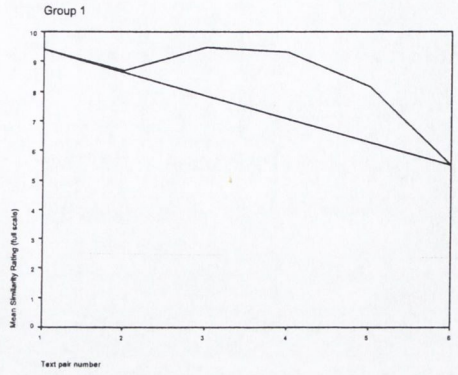


Figure 7.7: Mean Similarity Ratings for Group 1

7.5.2 Deviation from the central tendency

As an initial approach to the question of inter-rater agreement, an analysis was carried out to assess the degree of deviation of each rating from each respondent from two measures of central tendency: the mean similarity rating and the mode.[11] This analysis was initially carried out in order to identify outliers from among the respondents, i.e. respondents who consistently assigned ratings that deviated widely from the central tendency. The following analysis illustrates why none of the more deviant cases in the data were ignored as outliers.

Deviation of users from the mean rating

The range of values by which story pair ratings deviated from the average similarity rating for that story pair was from 0 to 6.38, a range of 6.38 points

[11] Appendix E.5.3 sets out a story-by-story matrix where the cells of the matrix are the average similarity rating. Appendix E.5.4 sets out the same matrix for the mode or most common rating given for a story pair.

on the original 11-point similarity scale. The average difference between a rating and the mean similarity rating for a given story was 1.4959. Analysis of the deviation of particular users from average similarity ratings reveals that the range of values by which respondents deviated from the average similarity rating over the six stories they rated was 0.22 - 4.39, a range of 4.17. On average, users disagreed with the mean similarity rating by 1.48 points on the original 11-point similarity scale with a standard deviation of 0.67. This illustrates that, although some respondents disagreed largely with the average rating, in general, they only strayed within approximately 2 points of the mean. This would suggest that respondents differ in their fine-grained distinctions of similarity but not in the ratings of similarity on a more coarse-grained scale.

The scatterplot in figure 7.8 shows that overall there is no correlation between the average rating assigned by a respondent, which might indicate a respondent bias, and the degree to which they deviated from the mean ratings. The most deviant respondents, however, were those who were biased towards low ratings, represented by the three points in the south-eastern quadrant of the scatterplot. As the overall tendency was towards high ratings, it is no surprise that the few low-biased respondents could be outliers in the data. It would appear that these three respondents may have interpreted the scale somewhat differently to the 319 other participants.

Deviation of users from the mode

Respondents' deviation from the mode gives a slightly different picture. The range of values is far wider, from 0 to 8 points on the similarity scale. The mean deviation from the mode, however, is only 1.39 with a standard de-

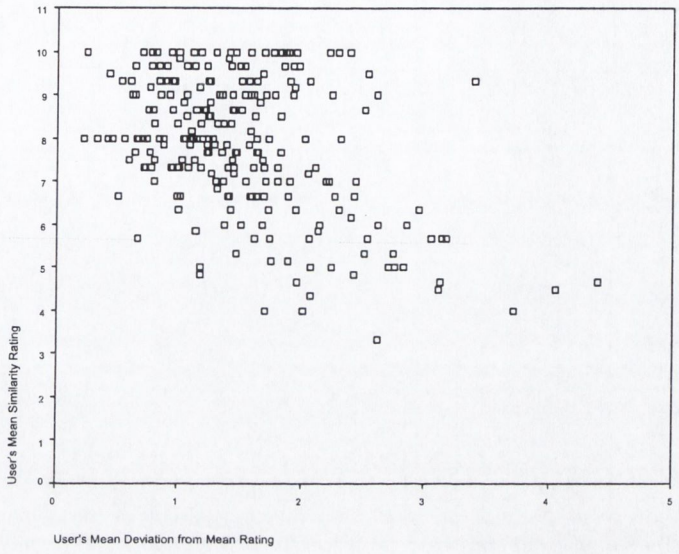


Figure 7.8: Scatterplot of mean similarity rating with mean user deviation

viation of 1.8. The nature of the mode as a measure of central tendency,[12] would dictate that many ratings, in this case 53% of ratings, are the same as the mode value. Over 80%, however, are within 2 points on the 11-point similarity scale from the mode. The same observation can be made here as in section 7.5.2 that this would indicate agreement on coarse-grained but not fine-grained similarity distinctions. Respondents' average deviation from the mode value varied within the range 0–5.33 with an average value of 1.363 and a standard deviation of 1.094. Again, respondents with a bias for lower ratings, for example, respondents 195, 147 and 155, accounted for the higher average deviation values. Figure 7.9 shows that the majority, 93.5% in fact,

[12]The mode in a dataset is the most frequently occurring value. In this case, the most frequently used rating per story pair.

of respondents, on average were within 3 points of the mode value.

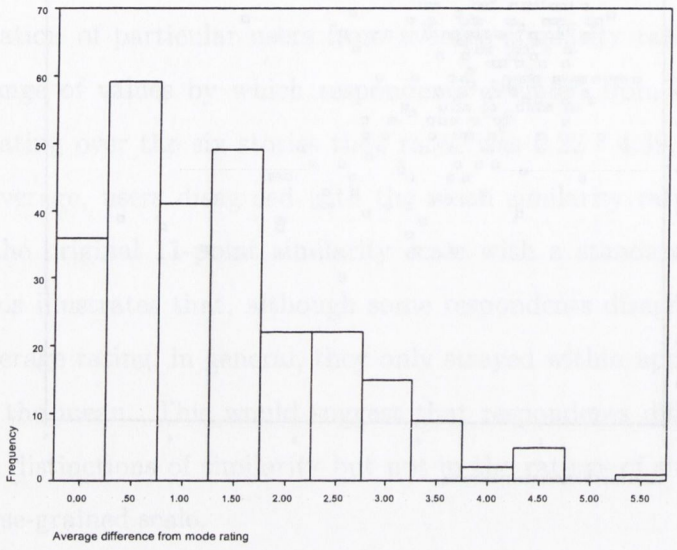


Figure 7.9: Histogram of average respondent difference from mode

Deviation of story pairs from the central tendency

A second analysis of deviations from central tendency measures was undertaken to identify which story pairs proved the most difficult to classify or, at least, which story pairs gave rise to the widest variation in user ratings. The range of deviation values in general are, of course, the same as above: 0-6.38 for deviation from the mean, 0-8 for deviation from the mode. The average deviation of each story from the mean and mode range from 0-3.6 and 0-2.89 respectively. On average, story pair ratings did not vary greatly. The mean deviation from the mean was 1.5 and from the mode 1.4 those with a standard deviation of 0.67 and 0.55 respectively. Again, the story pairs assigned a wider range of similarity ratings, with a greater average deviation,

tended to be those assigned lower ratings. This observation regarding variation in responses correlated with lower similarity ratings will be developed in section 7.5.5.

Some Comments

It would appear that the overall deviation from a central trend in the data is not high, on average under 2 points on the original 11-point scale which would reduce to 1 point on the 6-point similarity scale. No single individual was identified to exclude from further analysis as this would entail ignoring all low ratings. Nor was any story pair identified to exclude from further analysis as on average story ratings appeared to be quite consistent. These initial comments might suggest that the results of statistical tests for agreement would indicate a high level of agreement among respondents, given the lack of wide variation in the ratings given. The findings for the inter-rater agreement measures that follow, however, would appear to negate this assumption, although there is evidence to support the hypothesis put forward above that raters agree on coarse-grained similarity distinctions.

7.5.3 Percent Agreement

The first measure of inter-rater agreement I will discuss is percent agreement. Gale *et al.* (1992) measure inter-rater agreement on a word disambiguation task with a technique derived from Shipstone's Agreement-Disagreement ratio (Shipstone 1960). They term this measure Percent Agreement and it is calculated as follows:

$$\frac{\textit{number of observed agreements with the majority opinion}}{\textit{total number of possible agreements with the majority opinion}} \tag{7.7}$$

This measure was used extensively in empirical studies in the field of computational linguistics up to the late 90s. However, it has a significant shortcoming in that the measure does not account for unexpected agreement among respondents due to chance. Carletta (1996) examines the shortcomings of this and other measures of inter-rater agreement and concludes that, as such measures are not adjusted for chance, they cannot determine statistical significance of agreement and are not, therefore, readily interpretable. A second shortcoming of this test for these data in particular is that the measure only takes into account exact agreement of the ratings given which does not reflect the ordinal nature of the variable under analysis here.[13]

I report this measure, however, to give a first impression of agreement in the data and to provide a point of comparison between this study and other studies. The total number of possible agreements in the data is 1460. The observed number of agreements with majority opinion is 817.

$$\text{percent agreement value} = \frac{817}{1460}$$
$$\text{percent agreement value} = 55.96\%$$

The final percentage agreement measure of 55.96% is disappointingly low. In section 7.5.4, the κ coefficient statistic is calculated on the data to determine whether a chance-adjusted measure of agreement for ordinal data among participants is equally low.

7.5.4 Kappa Reliability Coefficient

The Kappa statistic is currently the *de facto* standard to measure agreement in coding tasks in computational linguistics and content analysis. Indeed,

[13]See section 7.5.7 for some further comment on the subject of determining agreement for ordinal variables.

the field of computational linguistics is indebted to Carletta (1996) for her analysis of agreement measures in use at the time and her exposition of κ as a viable, chance-corrected alternative to be adopted from the field of content analysis.

Currently, there is some variation in the calculation of the coefficient in use. Certain key aspects of this variation will be outlined below. The basic premise of all version of the test, however, derives from Cohen (1960). In informal terms, the basic premise is that the observed proportion of agreement for each rating assigned by raters is compared with the proportion of agreement for each rating which one could expect according to chance. The equation is as follows:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{7.8}$$

P(A) is the observed proportion of agreement in the data, P(E) is the proportion of agreement expected in the data according to chance. Observed agreement, therefore, is not evaluated in isolation but always in relation to agreement that might occur in similar data by chance.

To illustrate this premise, imagine two experiments, *ExperimentA* and *ExperimentB*. *ExperimentA* requires that respondents rate colours as dark or light. *ExperimentB* requires that respondents rate colours as blue, green, red or yellow. Common sense would have it that the odds of respondents agreeing on *ExperimentA*, a task requiring a binary judgment, are far higher than the odds of respondents agreeing on *ExperimentB*, a task requiring a selection from one of four categories. The odds of two raters agreeing on a rating in the two experiments are as follows:

- *ExperimentA* binary judgment odds: $0.5 * 0.5 = 0.25$
- *ExperimentB* 4-category judgment odds: $0.25 * 0.25 = 0.0625$

The P(E) coefficient in the calculation of κ is intended to account for these odds, i.e. for the degree of agreement that could be expected according to chance even in random data.

Interpreting the coefficient

The value of κ can range between -1 and 1. A κ coefficient of 1 indicates perfect agreement. A coefficient of -1 indicates perfect disagreement. Finally a coefficient of 0 indicates that raters agreed no more than was predictable by chance. In the event that $\kappa = 0$, observed agreement is no more than expected agreement, P(A)=P(E), and the numerator of the calculation is equal to 0.

Within the range [0,1], Brouwer *et al.* (1969) deem data to be reliable if $\kappa \geq 0.8$. They further adopt the position that tentative conclusions may be drawn as to the reliability of data where $0.67 < \kappa < 0.8$. These cut-off points are widely used in the field of computational linguistics. Rietveld and Van Hout (1993), however, are less stringent in their requirements. A κ coefficient in the range [0.2,0.4] is deemed to indicate fair agreement. In the psychiatric community, acceptable agreement is set at $0.5 < \kappa < 0.6$ (Grove *et al.* 1981). Grove *et al.* (1981) further note that not even the upper bound of 1 for Kappa in the population is fixed. Where other factors, such as specificity or sensitivity, in a population vary, the upper bound of Kappa can be well below 1.

Given the subjectivity of interpretation of κ, it is perhaps not surprising then that its use for measuring inter-coder reliability does come into question. It has been suggested that this statistic is too difficult to interpret and hence over-used in the literature (Uebersax 2004).[14] On the other hand, it does

[14]See (Uebersax 2004) for a discussion of the de-merits of the Kappa statistic as opposed

provide a standard used across the sciences and, as such, can be used to compare levels of agreement in different data. Researchers may then set the point of acceptability on the scale as appropriate to their field and the experiment in question.

Calculating P(E)

The value of P(E), the expected agreement, is central to the calculation of the κ coefficient. Consequently how to estimate P(E) is one of the main areas of discussion in the literature on the subject. In the two sample experiments set out above, each rating was deemed equally likely and hence the probability of 1 was divided equally between the possible ratings. In real experiments, however, as it is very unlikely that the distribution of different ratings in the population as a whole is known, P(E) must be estimated from the distribution of the ratings in the data.

There is a fundamental divide between those who assume the distribution of proportions is the same for all coders, (Krippendorff 1980, Siegel and Castellan 1988) and those who do not, (Cohen 1960). Di Eugenio and Glass (2004) provide an in–depth analysis of how this assumption impacts on the final value of κ. The data of the current study has a maximum of 10 ratings per individual. In section 7.5.2 we saw that the respondents generally did not vary too widely from the central tendency. Therefore, it would seem appropriate to calculate the expected agreement value under the assumption of equal distribution of proportions among raters. The calculation of the coefficient according to Krippendorff is somewhat different as it is framed in terms of observed and expected *dis*agreement but the assumption of equal distribution among raters holds here also (Krippendorff 1980, p.138).

---

to latent trait analysis.

The two sections that follow set out the details of the calculation of the κ coefficient on the data according to Siegel and Castellan (Siegel and Castellan 1988) and according to Krippendorff (Krippendorff 1980).

Siegel and Castellan's κ coefficient

The equation to calculate κ is as follows:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{7.9}$$

The calculation can be broken down into counting the observed agreement, P(A), and estimating the expected agreement, P(E).

P(E) Where k raters rate N story pairs, the proportion of story pairs assigned some rating j (p_j) is the total number of story pairs assigned rating j (C_j) divided by the total number of ratings (Nk).

$$p_j = \frac{C_j}{Nk} \tag{7.10}$$

The expected agreement for each category, then, is the proportion of ratings assigned to this category squared (p_j^2). According to Siegel and Castellan, P(E) is simply the sum of all squared proportions values over each of the categories available for rating, as in equation 7.11.

$$P(E) = \sum_{j=1}^{m} p_j^2 \tag{7.11}$$

P(A) For each story pair, the degree of agreement among the ratings of that story pair is the proportion of paired ratings which agree to the proportion of possible pairs of ratings. The formula for S_i, agreement on story pair i, is set out in equation 7.12. In this formula, k is the number of raters, m is the number of possible ratings and n_{ij} is the number of times rating j was given

for story pair i. Where there is no agreement on a particular rating j, i.e. when one or no rater assigned rating j, i.e. $n_{ij} = 1$ or $n_{ij} = 0$, that rating adds nothing to the agreement sum, i.e. $n_{ij}(n_{ij} - 1) = 0$. This formula, then, calculates agreement between all k raters on all m ratings for story pair i.

$$S_i = \frac{1}{k(k-1)} \sum_{j=1}^{m} n_{ij}(n_{ij} - 1) \tag{7.12}$$

P(A), the total proportion of agreement, is the average of all S_i proportions. The formula is set out in equation 7.13 where N is the number of story pairs.

$$P(A) = \frac{1}{N} \sum_{i=1}^{N} S_i \tag{7.13}$$

The table in appendix E.3 sets out the proportions used to calculate Siegel and Castellan's (1988) κ coefficient on the data for the experiment reported here. The table is a 192-by-6 matrix where the columns are the 6 ratings of the reduced similarity scale and the rows are each of the story pairs rated. Each row sets out the rating counts for each story pair i. S_i, the proportion of agreement for each story pair, is reported in the last column of each row. C_j, the total number of story pairs assigned a particular rating j, is listed in the penultimate row. The final row lists p_j, the proportion of ratings assigned to category j.

Kappa P(A), then, is the sum of all S_i values and P(E) is the sum of all squared p_j values. For the experiment reported here, this works out as follows:

$$P(A) = 0.3549$$
$$P(E) = 0.3024$$
$$\kappa = \frac{0.3549 - 0.3024}{1 - 0.3024}$$
$$\kappa = 0.07522$$

This value for κ strongly suggests that there is no more consensus in the data than predicted by chance. However, this calculation does not take account of the fact that the ratings are on an ordinal scale. A rating of 0 is taken to be equally different from a rating of 1 or of 6. Krippendorff (1980, p.138) states that his measure of agreement, Krippendorff's α, can be generalised to multiple raters and multiple rating categories and multiple types of data whether ordinal or canonical. The following section details the calculation of Krippendorff's α.[15]

Krippendorff's α

Krippendorff approaches the question of agreement from the opposite perspective, looking at degree of disagreement, observed and expected. The basic equation is as follows:

$$\alpha = 1 - \frac{observed\ disagreement}{expected\ disagreement} \tag{7.14}$$

A reliability data matrix is constructed of the data. For the experiment reported here, this is a 192-by-10 matrix of 192 story pairs by $m = 10$ respondents. Each $cell_{ij}$ in the matrix contains the rating given by rater i for story pair j. As not all story pairs have 10 ratings, some cells will be empty. Therefore, m_u, the number of values in each column, will vary. The first six columns of the matrix, the data for the first six story pairs, are set out in table 7.3

The coincidences, instances of pairs of ratings, within story pairs are then tabulated as a coincidence matrix. For this experiment which uses a 6-point rating scale, this is a 6-by-6 matrix where each $cell_{ij}$ is a count of the times rating i and j co-occurred in a story pair. Each co-occurrence is entered

[15]Krippendorff uses α to represent the κ coefficient of agreement.

	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6
Rater1	4	4	5	5	5	3
Rater2	-	-	-	-	-	-
Rater3	5	5	5	5	5	4
Rater4	5	4	-	-	-	-
Rater5	-	-	-	-	-	-
Rater6	5	5	5	5	4	1
Rater7	4	4	3	4	3	3
Rater8	5	5	5	5	4	4
Rater9	5	3	5	4	3	1
Rater10	-	-	-	-	-	-

Table 7.3: Reliability Data Matrix of 6 story pairs (6-point similarity scale)

twice, once at $cell_{ij}$ and once at $cell_{ji}$. In table 7.3, the Pair1 column contains ten 4-5 co-occurrences, ten 5-4 co-occurrences, two 4-4 co-occurrences and twenty 5-5 pairs to give a total of 42 co-occurrences. Each co-occurrence in column u adds $\frac{1}{(m_u-1)}$ to the appropriate cell in the matrix, where m_u is the number of ratings in column u. This takes account of any missing ratings in the data set. The coincidence matrix for the experiment data is set out in table 7.4

Krippendorff states:

> Note that α accounts for different metrics of levels of measurement by weighing the observed (and expected) coincidences by the squared difference between the coinciding values. Differences can be expressed as mathematical functions and in the form of a

	1	2	3	4	5	6	Total
1	**0.00**	0.71	0.94	0.43	0.17	0.37	2.62
2	0.71	**9.40**	10.92	9.96	8.78	11.41	51.18
3	0.94	10.92	**19.91**	25.70	30.90	49.10	137.47
4	0.43	9.96	25.70	**35.80**	48.30	73.60	193.79
5	0.17	8.78	30.90	48.30	**81.09**	135.36	304.6
6	0.37	11.41	49.10	73.60	135.36	**311.56**	581.4

Table 7.4: Coincidence Matrix for calculation of Kappa Coefficient

table. ... Ordinal differences depend on their frequencies of using values. (Krippendorff 2004, from his α coefficient website)

This experiment data is ordinal where each value is a rank and the difference between ranks depends on how far apart they are. Table 7.5 sets out the ordinal metric differences, ordinal δ^2_{ck}, for the experiment data. The formula to calculate the difference metric is set out in equation 7.15.

$$ordinal\ \delta^2_{ck} = \left(\frac{n_c}{2} + \sum_{g>c}^{g<k} + \frac{n_k}{2} \right) where\ c < k \tag{7.15}$$

This metric should account for the fact that in the data a rating of 4 is less different from a rating of 5 than it is from a rating of 0.

The α coefficient then is calculated from the values in the coincidence matrix and the ordinal metric differences matrix according to the following formula, where D_o is the observed disagreement and D_e is the expected disagreement:

$$\alpha = 1 - \frac{D_o}{D_e}$$

	1	2	3	4	5	6	
1	0.00	723.01	14690.83	82272.34	287313.44	958444.11	n_1=2.62
2	26.89^2	0.00	8895.67	67570.19	259210.69	906518.59	n_2=51.18
3	121.21^2	94.32^2	0.00	27431.90	172067.76	735813.69	n_3=137.47
4	286.83^2	259.94^2	165.63^2	0.00	62093.03	479099.36	n_4=193.79
5	536.02^2	509.13^2	414.81^2	249.18^2	0.00	196235.99	n_5=304.6
6	979.00^2	952.11^2	857.80^2	692.17^2	442.99^2	0.00	n_6=581.4

Table 7.5: Matrix of Difference Metrics for calculation of Kappa Coefficient

$$= 1 - (n-1)\frac{\sum_c \sum_{k>c} obs_{ck}\ metric\ \delta^2_{ck}}{\sum_c \sum_{k>c} n_c * n_k\ metric\ \delta^2_{ck}}$$

The numerator is the sum of observed disagreement multiplied by the appropriate difference metric. That is, the sum of the products of $cell_{ij}$ of the coincidence table 7.4 and $cell_{ij}$ of difference metric table 7.5 where $i \neq j$ and $i < j$. This formula then is only concerned with the values in one of the off-diagonal triangles of the symmetrical matrices. The denominator is the sum of expected disagreement multiplied by the appropriate difference metric. That is, the sum of the products of the total of row i, the total of row j and $cell_{ij}$ of difference metric table 7.5. The result of the calculations for these data is:

$$\alpha = 0.2034$$

This value is just beyond the threshold of fair agreement according to Rietveld and Van Hout (1993) but well below the 0.67 value generally accepted as the baseline for tentative conclusions of agreement.

As a further step, for both versions of the kappa statistic, a random data

set was generated for the same rating categories and the statistic calculated on the new data set. The values were as follows:

- Siegel and Castellan: $\kappa = -0.0006$
- Krippendorff: $\alpha = .0102$

In both cases, the real data gave a much higher statistic than the random data. However, random number generators, while claiming to be random, generally generate numbers according to a normal distribution. The experiment data is very far from normal, perhaps coming closest to a binomial distribution. It would therefore be unwise to draw conclusions from this as to the data being more consistent than random data, given they are not distributed in the same manner.

At this point, one might conclude that agreement among respondents as to the similarity of story pairs was low. Indeed, further testing could provoke criticism for generated results and possibly unreliable conclusions only by repeated testing. Through repeated testing and, therefore, re-processing, of experiment data, it is possible that the data may end up conforming to some expected standard. However, in this instance, it is not a question of re-processing data but rather of determining which statistical test is appropriate for these data. To this end, section 7.5.5 re-examines the experiment data and investigates the degree of agreement in more detail. In response to the analysis, a further agreement measure, the Cochran Q statistic is applied to the data.

7.5.5 Re-analysing the data

Correlation of respondent agreement with mean ratings

A regression analysis

On observation of the agreement scores for each story pair, it appeared that the lower the similarity ratings overall for a story pair, the less agreement among the participants. It would seem that respondents agreed readily on story pairs they found to be dissimilar but where respondents noted some similarity in the stories, they failed to agree on the *degree* of similarity. This observation was borne out by a regression analysis of percentage agreement against mean rating value for story pairs.

Figure 7.10 shows a scatterplot of mean rating value against percentage agreement for story pairs. A single data point can be observed in the upper left hand corner of the plot. This single case which has a high percentage agreement value and a very low mean similarity rating stands out as an outlier in the data. Figure 7.11 shows the plot of the data with this outlier removed and with a best-fit line overlaid on the data. The plot suggests a positive correlation between mean similarity rating and percentage agreement of respondents. Furthermore, as both mean rating and percentage agreement values have a limited range, 0 to 10 and 0 to 1 respectively, it is important to note that the fitted line does not predict values outside the range of one variable for values within the range of the other variable.

A regression analysis of the data was carried out. The results are set out in the tables below. A linear regression model was produced for percentage agreement among respondents based on mean similarity rating. Table 7.6 shows the model coefficients. The ANOVA table 7.7 indicates the acceptability of the model statistically. The sum of squares values indicate that the

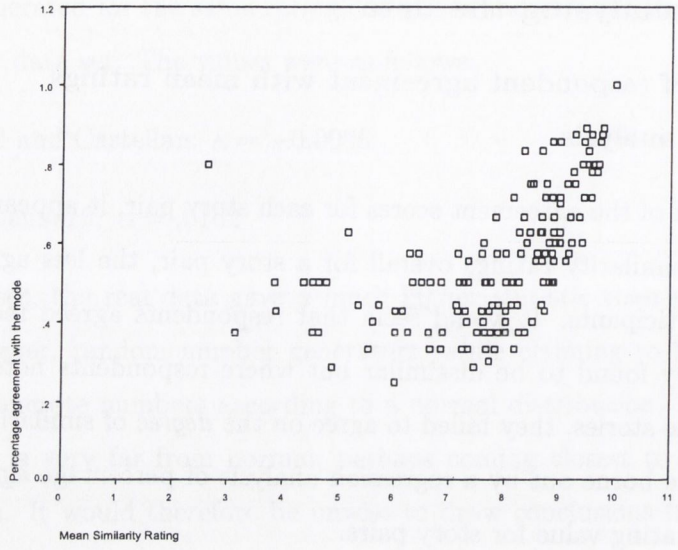


Figure 7.10: Scatterplot of percentage agreement with mean similarity rating

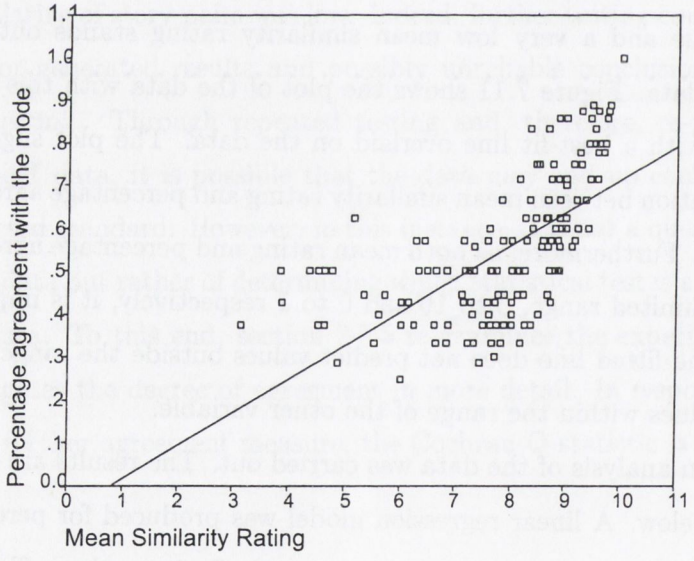


Figure 7.11: Scatterplot of percentage agreement with mean similarity rating without outlier and with fitted line

model accounts for approximately 40% of the data. The significance value of the F score is $p < 0.001$ which would strongly suggest that this variance in the data for which the model accounts is not due to chance.

There is strong evidence, therefore, to suggest that there is some relationship between the predictor, mean similarity rating, and the dependent variable, percentage agreement. The model summary table 7.8, then, reports the strength of that relationship. The correlation coefficient R is large at 0.638, indicating a strong relationship. The coefficient of determination, R Square, is 0.407. The model therefore explains approximately 40% of the variance in the data and the linear relationship between the dependent variable, percent agreement, and the model derived from the predictor, mean similarity rating, is quite a strong one. A comparison of the standard deviation of percent agreement (see table 7.9) with the standard error of the model estimates gives a further indication of the strength of the model fit. The standard error of the model (12.52835) is considerably lower than the standard deviation of the dependent variable (16.22645). Hence, this model is a more effective predictor of a percentage agreement value than a measure of central tendency of the variable proved to be.

Model	Unstandardised Coefficients		t	Sig
	B	Std Error		
1 (Constant)	-6.024	5.494	-1.096	.274
Mean similarity rating	7.743	.680	11.390	.000

Table 7.6: Regression Coefficients

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	20361.240	1	20361.24	129.723	.000
Residual	29665.347	189	156.960		
Total	50026.587	190			

Table 7.7: ANOVA, Dependent Variable: Percentage Agreement, Predictors: (Constant) Mean Similarity Rating

Model	R	R Square	Adjusted R Square	Std. Error of Estimate
1	.638	.407	.404	12.52835

Table 7.8: Linear Regression Model Summary Table

	Mean	Std. Deviation	N
Percentage agreement	55.6952	16.22645	191
Mean similarity rating	7.9708	1.33693	191

Table 7.9: Descriptive Statistics

Correlation of respondent agreement with mean ratings
Curve estimation analysis

The linear model discussed above (7.5.5), although effective, is perhaps not the best model. A second look at the scatterplots in figures 7.10 and 7.11 would suggest a quadratic model might be more appropriate to the data. A curve estimation analysis was carried out on the data and, in fact, a quadratic model provides a better fit for the data. The model summary table 7.10 shows a higher R Square value (0.63125) indicating that the model accounts for approximately 63% of variance in the data. Furthermore, the standard error of the estimate is much lower.

Model	R	R Square	Adjusted R Square	Std. Error of Estimate
1	.79451	.63125	.62733	.09906

Table 7.10: Curve Estimate Model Summary Table

Implications of the Regression Analysis

The linear regression and curve estimation models lend support to the observation made in section 7.5.2 that respondents tended to agree more readily on story pairs they deemed dissimilar and tended to disagree more on story pairs that they deemed to be similar to some degree. It would appear, given the better fit of the curve estimation model, that the relationship between agreement and degree of dissimilarity is complex and probably not linear but the tendency for agreement to increase as the degree of dissimilarity increases is borne out in both models. This hypothesis, therefore, raises two issues.

The first issue regards the experiment methodology. Perhaps, in hindsight, magnitude estimation would have been a more appropriate scoring

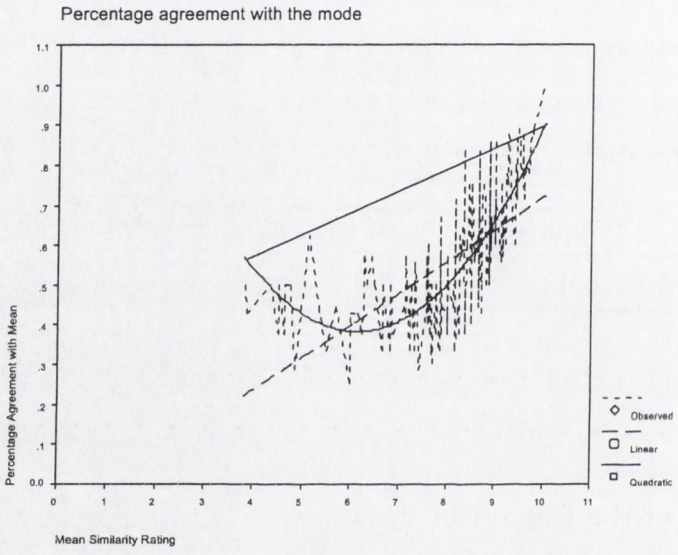


Figure 7.12: Plot of Observed Values, Linear and Quadratic Curve Estimation Models

metric to use. The lack of agreement among participants on degree of similarity may be a by-product of the scale available to them rather than an indication of real disagreement. The more flexible rating mechanism provided by magnitude estimation could have forestalled this problem.

The second issue regards the tests for inter-respondent agreement appropriate to this study. In previous sections, the results of the standard inter-respondent agreement measures were outlined. The measures indicated that agreement among respondents was low. However, if, as it would appear, the level of agreement is dependent on the mean degree of similarity of two text pairs, it may be appropriate to judge agreement by a different measure.

The measure I suggest is the Cochran Q statistic. This measure is generally used as an agreement metric for dichotomous data. As respondents appear to make a binary decision between dissimilar and somewhat similar story pairs, a transformation would seem appropriate from the scaled judgements in the study to judgments making the binary distinction between text pairs that they deem completely dissimilar *versus* those with some similarity. The results of this analysis are set out in section 7.5.6.

7.5.6 Cochran Q statistic

The Cochran Q test (Cochran 1950) tests the hypothesis that several dichotomous variables have the same mean. In this context, the null hypothesis of the Cochran Q test is that any agreement between respondents (on a dichotomous task) is due to chance. This test only works for binary variables, hence, as suggested in section 7.5.5, the data from the study was transformed so that all ratings for "completely different" were given the value 1 and all other ratings, ratings that were deemed to have some degree of similarity, were given the value 0.

As this test requires that each variable has the same number of observations (i.e. each story pair have that same number of ratings), the Cochran Q statistic was calculated on 6 different subsets of the data. The subsets were determined according to the number of ratings that were given for each text pair. The results are set out in table 7.11. In all but the first case where there are only 5 ratings per text pair, the null hypothesis that any agreement is due to chance may be rejected at $p < 0.01$. These tests would strongly suggest that respondent agreement, at least on a binary task of assigning the label "completely dissimilar" or "somewhat similar" to a story pair, is statistically significant.

7.5.7 Euclidean distance in 6-dimensional space

The final measure to evaluate inter-rater agreement that I will discuss is Euclidean distance. For each respondent that completed all six story pair evaluations, the Euclidean distance between that respondent and all other respondents in the group to which they belonged was calculated. This measure is not given as a statistical test but rather the results are presented visually to give a sense of consistency, or lack thereof, within the groups of raters. Euclidean distance is ideally suited to measure differences in a rank-ordered variable, such as this similarity rating, as inherent to the calculation is the notion that a rank of 4 and a rank of 5 are closer than a rank of 4 and a rank of 1. This characteristic of ordered nominal data is provided as a special case of the Kappa statistic but is not a given for Kappa or other agreement metrics.

Each respondent gave a total of six ratings. Each rating is taken as a co-ordinate for a particular dimension in 6-dimensional space: story pair 1 is dimension1, story pair 2 is dimension 2, etc. Hence, each respondent is

# Ratings per story pair	Test Statistic	
5 ratings	N	5
	Cochran's Q	12.100
	degrees of freedom	11
	Asymp. Sig.	.356
6 ratings	N	6
	Cochran's Q	52.376
	degrees of freedom	28
	Asymp. Sig.	.003
7 ratings	N	7
	Cochran's Q	91.850
	degrees of freedom	49
	Asymp. Sig.	.000
8 ratings	N	8
	Cochran's Q	108.144
	degrees of freedom	45
	Asymp. Sig.	.000
9 ratings	N	9
	Cochran's Q	66.823
	degrees of freedom	38
	Asymp. Sig.	.002
10 ratings	N	10
	Cochran's Q	33.919
	degrees of freedom	15
	Asymp. Sig.	.003

Table 7.11: Cochran Q statistics for 6 data subsets

represented as a point in 6 dimensional space defined by the ratings they gave for each story pair. Every group contained ten respondents but only those respondents that completed the experiment were included in this distance evaluation.

The Euclidean distance between respondents was calculated in Matlab with the *pdist* function. $Y = pdist(X)$ computes the Euclidean distance between pairs of objects in an $m - by - n$ matrix X, which is treated as m vectors of size n. For a data set made up of m objects, there are $\frac{m(m-1)}{2}$ pairs. The output, Y, is a vector of length $\frac{m(m-1)}{2}$, containing the distance information. The distances are arranged in the order (1,2), (1,3), ..., (1,m), (2,3), ..., (2,m), ..., ..., (m-1,m). Y is also commonly known as a similarity matrix or dissimilarity matrix. Given an $m - by - n$ data matrix X, which is treated as m 1-by-n row vectors $x1$, $x2$, ..., xm, the various distances, d_{rs} between vector xr and xs are defined according to equation 7.16.

$$d_{rs} = \sqrt{(x_r - x_s)(x_r - x_s)'} \tag{7.16}$$

The data for each group was presented as an $m - by - 6$ matrix where m is the number of respondents in the group that completed the experiment. The output was an $m - by - (m - 1)$ matrix of the distances between each respondent, each row of the input matrix. The values along the right-left diagonal of this output matrix are all 0: each rater is at 0 distance from themselves. Furthermore, the matrix is symmetrical about this diagonal: the difference between raters x and y is the same as the difference between raters y and x. The output matrix was then rendered as a colour m-by-(m-1) graph where the square (x, y) in the grid represents the distance between respondent x and respondent y. The graph also is symmetrical about the diagonal. The colour graph and the input and output matrices for each group are set out in Appendix E.4.

The first three data sets in the appendix are artificial data given as a point of reference in order to assess the colour scale on the graphs of the real data. Graphs E.66 and E.67 illustrate the colour pattern in the case of complete agreement among raters, all raters gave the same rating for each story pair. The graph is a uniform dark blue, Euclidean distance of 0. The former graph derives from similarity ratings of 1 and the latter from similarity ratings of 10. It is apparent, then, that the Euclidean distance value is unaffected by the actual rating given and is simply a function of the *difference* between the ratings of individual raters.

Graph E.65 illustrates the case of total disagreement among raters. In the input matrix, each rater has assigned only one value and that value is different for each rater, ranging from 1 to 10. The dark blue right-left diagonal illustrates that each rater is in perfect agreement with themselves, Euclidean distance of 0. The red squares represent the maximum difference, Euclidean distance of 22, between rater 1 and rater 10 in this hypothetical group. The spectrum of colour from blue to red illustrate the range of possible differences between raters according to the ratings they assigned. For example, rater 2 and rater 3 assigned close ratings of 2 and 3 respectively for all story pairs. The distance of 2 between the two raters is represented by a light blue tone. Whereas, the much larger distance measure of 15 for rater 2 and rater 8, who assigned ratings of 2 and 8 respectively, is represented by an orange colour.

The colour scale of graph E.65 is used for all of the graphs derived from the real group data. In all the graphs, then, dark blue indicates perfect agreement (the right-left diagonal in all graphs) and red indicates complete disagreement. The intervening spectrum represents degrees of agreement from almost perfect agreement to almost complete disagreement. I stress that this is not a test of statistical significance but the aim is rather to

visualise agreement trends across the data.

In the graphs of the thirty-two groups, medium-to-light blue predominates[16] which would suggest that raters assigned ratings which tended to be closer rather than distant on the similarity scale. Graph E.69 is a prime example of this. The graphs also highlight respondents which may be considered as outliers. Rater 5 in group 20 is such a case. Indeed, this respondent is number 195 identified as a respondent with low-bias in section 7.5.2. This individual disagreed substantially with all other respondents in the group. As such, there is a marked yellow cross in graph E.87, reprinted here as figure 7.13, in contrast with the dark blue of the remaining squares of the grid.

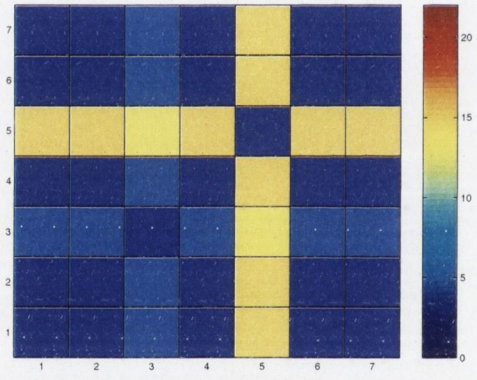


Figure 7.13: Group 20 Euclidean Distance Matrix

Graph E.90 for group 23, reprinted in figure 7.14, illustrates a different case, where overall respondents were in agreement, with the exception, however, of raters 2 and 5. Rater 5 in this group was biased towards very low ratings while Rater 2 tended towards the higher end of the scale. These extreme points of view within the group are illustrated by the single orange squares in a mass of blue-to-light-green.

[16]The average Euclidean distance measure was 5.36 with a standard deviation of 2.81.

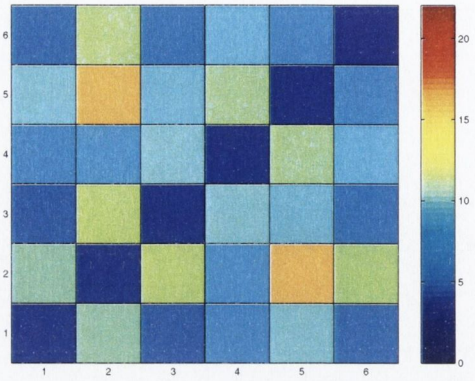


Figure 7.14: Group 23 Euclidean Distance Matrix

Furthermore, where fewer respondents in a group completed the experiment, for example graph E.86 and E.93, the Euclidean distances would appear to rise. This reflects the results of the Cochran Q statistic reported in section 7.5.6 where only stories rated by less than 6 respondents yielded a non-significant result for the null hypothesis that agreement between respondents was due to chance. This would suggest that a data set of five raters per story pair is not sufficient to determine rating agreement for this task.

Although this graphical representation of Euclidean distance between raters does not give a definitive statistical test of agreement, it does support the observation made in section 7.5.5 that raters tend to agree on a basic similar-dissimilar distinction but not on the fine-grained degree of similarity measure. Overall, there appears to be a trend for raters to give close but not identical ratings for story pairs.

7.5.8 Reliability of the Training Data

In sections 7.5.6 and 7.5.7, I noted that the story pairs rated by under six raters did not give statistically significant results for rater agreement. In this section, I investigate agreement in the responses given on the experi-

ment training page. The 321 respondents all completed the training page and therefore, there is no issue with data sparsity or incomplete ratings in calculating an agreement metric.

The training page consisted of two parts:

1. a set of ten yes/no questions on different aspects of similarity or relatedness

2. a similarity rating scale and a relatedness rating scale, as in subsequent experiment pages

For the dichotomous data of part one of the training page, the Cochran Q statistic was used to test for agreement among respondents. The results are set out in table 7.12. The null hypothesis that agreement among respondents on the ten questions was due to chance may be rejected at the significance level $p < 0.001$.

	Test Statistic	
Training Data	N	299
	Cochran's Q	674.770
	degrees of freedom	9
	Asymp. Sig.	.000

Table 7.12: Cochran Q statistics for training data

The distribution of the similarity and relatedness ratings are set out in table 7.13 It is clear that the distribution of similarity and relatedness values for this story pair is normal. This would suggest that there is little agreement among respondents on the degree of similarity or relatedness of the two stories. The consensus on the dichotomous tasks contrasted with the

Rating	Similarity	Relatedness
Missing	1.2%	2.2%
0	0%	0.6
1	0%	0%
2	3.7%	3.4%
3	1.6%	2.2%
4	25.9%	20.9%
5	6.5%	7.2%
6	26.5%	24.6%
7	5.0%	4.4%
8	17.1%	15.0%
9	2.2%	0.9%
10	10.3%	18.7%

Table 7.13: Distribution of Similarity and Relatedness Ratings in Training Data

lack of consensus on the rating task might indicate that respondents had not settled in to using the ranked list at this point. The reliability results for the test data itself suggest that respondents never in fact came to terms with the ranking task but the binary task of stating there was some degree of similarity was one which they undertook easily from the start.

7.5.9 Conclusions

The agreement metrics set out in previous sections would seem to support two hypotheses.

1. Humans are consistent in identifying story pairs that are deemed completely different and also in identifying story pairs with some degree of similarity

2. Respondents did not agree on the degree of similarity in a story pair deemed not completely dissimilar

The lack of agreement on a degree of similarity may be an effect of the methodology of this study or it may relate to issues which cannot be controlled for, like respondents' world-knowledge. Whatever the source of the phenomenon, the implications for this research is that a gold standard of texts from the RTÉ news corpus may be used to test the system I developed and reported on in chapters 5 and 6 but only for a binary *no similarity — some similarity* distinction. The evaluation of system performance using this "silver standard" is set out in chapter 8.

The research discussed in this chapter has fulfilled three functions. Firstly, a gold standard of human similarity judgments has been compiled for 192 text pairs from the RTÉ news text corpus. This standard is used to evaluate the performance of the system I developed to automatically estimate text

similarity. Secondly, the discussion of the experiment design highlights some issues regarding text similarity and text relatedness which I will return to in chapter 8. Finally, this chapter provides a detailed encapsulation of the literature on measuring reliability, with my experiment as a case study. This analysis shows the various conclusions to which different statistical methods lead and highlights the importance of choosing a test that is appropriate to the experiment design and data, as Bishop *et al.* (1975) suggest in the citation in section 7.5.

Chapter 8

Results and Evaluation

8.1 Introduction

This chapter describes the results of the evaluation process for the system I designed and implemented which was described in chapters 5 and 6. The core component of the system is the similarity detection engine. Sections 8.2 to 8.5 set out an in-depth analysis of this component. As I noted in the previous chapter, the results of the gold standard experiment would suggest that humans are not consistent in evaluating the *degree* of similarity between texts. The results do suggest, however, that they are consistent in evaluating whether there is some similarity or none whatever between texts. For this reason, the evaluation discussed below is based on the binary task of detecting when texts have something in common. The results reported regard how accurately the system identifies those texts that humans rated as having something in common as well as its accuracy in identifying those texts that humans rated as completely dissimilar. Section 8.2 discusses the three gold standard datasets used to evaluate the system and the key differences between them. Section 8.3 discusses the process of parameter estimation undertaken to derive optimum thresholds for the three datasets. Section 8.4 discusses the metrics by which system performance has been evaluated. Section 8.5 sets out results for each data set. These are synthesised in section 8.5.4 in a discussion of the conditions for optimum system performance. The analysis of the similarity detection component closes with a discussion of some questions regarding the concept of text similarity which emerged over the course of the evaluation process. The second component of the system identifies differential content in similar text pairs. Section 8.6 describes a small scale evaluation of this component on a corpus of "Google News" texts.

8.2 Three "Gold" Standard Datasets

8.2.1 RTÉ Data

Chapter 7 sets out the details of an experiment to define a gold standard of human similarity judgments on a selection of texts from a corpus of RTÉ texts. Details of the corpus compilation methodology and the corpus contents are described in section 3.3.1. In brief recapitulation, the corpus consists of short news texts collected by the researcher from the Irish National Broadcaster, RTÉ. The texts include source news wire texts, broadcast news scripts and articles from the broadcaster's news website (http://www.rte.ie/news) on a variety of topics.[1] A subset of the corpus was used in the Gold Standard experiment resulting in 6 to 10 similarity judgments for each of 192 text pairs where respondents assigned a similarity ranking for each text pair in response to the question:

> How similar are these texts?

The first Gold Standard dataset for this system performance evaluation uses the mean similarity judgment for each text pair.[2]

8.2.2 RTÉ Rel Data

A second Gold Standard dataset is also derived from the RTÉ experiment described in chapter 7. However, the Gold Standard text pair judgment in this case is the mean of the responses for each text pair to the question:

> How related are these texts?

[1] See Appendix A.1 for a list of story topics.

[2] See chapter 7 for a full discussion of the experiment data and inter-rater agreement results.

The task of making the relatedness judgment was presented in the same way as and simultaneously with the similarity judgment task. The motivation for including this dataset is set out in section 8.2.4.

8.2.3 Australian Data

The third Gold Standard dataset was very kindly donated by Dr. Michael Lee at the University of Adelaide, Australia. The experiment described in chapter 7 had much in common with an experiment carried out by a group of researchers at this University (Lee *et al.* 2004). The corpus consists of 50 documents from the Australian Broadcasting Corporation news mail service which provides e-mails of the headline news stories of the day.

Australian Similarity Experiment

The similarity judgment experiment was carried out on all pair-wise combinations of these 50 texts, giving a total of 1225 text pairs. The subjects were a group of 83 students at the University who were paid ten dollars for every 100 text pairs rated. Each of the 1225 text pairs was rated 8–12 times.[3] Ratings were given on a 5-point scale, from 1 (highly unrelated) to 5 (highly related. The authors report a distribution of ratings which is similar to the distribution of the RTÉ experiment results described in section 7.4.2. They also report high inter-rater agreement for similarity ratings.

8.2.4 Australian and RTÉ Experiment Differences

While similar in many ways, the Australian and the RTÉ experiments have several points of contrast that should be discussed.

[3]Due to an error in their experiment presentation software, the original intention of gathering exactly ten ratings per document pair was not fulfilled.

1. **Data size.** The RTÉ dataset is approximately 6.5 times smaller than the Australian. This has implications for threshold estimation (section 8.3) and performance evaluation (section 8.5) where results on the Australian data carry more weight given the size of the dataset.

2. **Number of ratings per respondent.** Australian respondents carried out on average 147 ratings each and were paid a nominal fee for these ratings. One might expect that the Australian respondents would become very skilled at the task of judging similarity over the course of the 147 ratings.[4] There is also the added incentive of payment to promote "good respondent" behaviour.[5] These two factors could combine to provide consistency in results which would be reflected in the reported high inter-rater agreement. RTÉ respondents, on the other hand, were solicited from among friends and colleagues and were asked to complete the training page and six other text pair ratings. The personal connection to the researcher could also promote a strong drive for "good respondent" behaviour. However, the low number of ratings required could lead to uneven, in a sense, untrained, results.

3. **Rating Task.** A third significant difference lies in what exactly the subjects were asked to rate. In the Australian experiment, they were posed the question:

   How similar is the **topic** of these two articles? (My emphasis)

[4]It would be an interesting analysis to compare inter-rater agreement on raters' initial ratings with agreement on raters' final ratings to investigate if training had a positive effect on agreement scores.

[5]Of course, "good respondent" behaviour is considered an unfortunate confounding factor in many psychological experiments (see section 7.4.1).

This question gives a very precise focus for determining text similarity. In contrast, during the talk-aloud protocol for the RTÉ experiment (section 7.3.1), the two subjects commented on the difficulty of defining a concept of text similarity. This is due to the fact that texts may be similar in many ways and on many levels. This prompted the introduction of a training page for the RTÉ experiment in which respondents are encouraged to address different aspects of text similarity and give a final rating based on an overview of these aspects. The concept of text similarity in the RTÉ experiment was, therefore, much broader and more difficult to define relative to the Australian experiment.

4. **Rating scale.** A related difference regards how the rating scale was expressed. The 5-point rating scale of the Australian experiment was expressed in terms of relatedness, "one indicating 'highly unrelated' and five indicating 'highly related"'(Lee *et al.* 2004, p.2). The RTÉ rating scale for the similarity question, on the other hand, was expressed in terms of similarity, 0 indicating identical and 10 completely dissimilar. The issue of a relatedness—similarity distinction also arose in the talk-aloud protocol. The two respondents noted that relatedness of topic could confuse respondents perceptions of text similarity. For example, a comparison of a text that reports a crime and a text that reports the subsequent court case for that crime are clearly highly related. However, it does not necessarily follow that the two are similar. For this reason, the RTÉ experiment posed an additional question:

   Are these texts related?

and although close to the similarity results, the relatedness ratings displayed a pattern of their own that merits analysis in its own right.

The previous chapter noted that respondent feedback on the RTÉ similarity judgment task was focused on the difficulty of the task of evaluating similarity. The Australian research group, however, had no such feedback regarding difficulties encountered.[6] The differences noted above may go some way to providing an explanation for this as Australian respondents had more time to get used to the similarity rating task and the focus of the judgment was clearly defined as strictly delimited to topic area. This could also explain the reported high levels of inter-rater agreement in the Australian experiment relative to the RTÉ experiment.

The difference in task definition would suggest that the notion of "text similarity" is not clearly defined. A single-level concept of similarity as similarity of topic is valid but so too is a more general multi-level concept. This question provides an added dimension to the evaluation of system performance described in this chapter as the Australian data can be taken as representing topic similarity and RTÉ as a more general notion of similarity.

The difference in rating scale definition would suggest that there may be some uncertainty as to whether the Australian respondents were judging similarity or relatedness. The RTÉ and Australian Gold Standard datasets, therefore, could represent not just a mere shift in focus but a different task entirely. The RTÉ Rel Gold Standard dataset was included in the evaluation process to address this issue. The RTÉ experiment separated out the judgment of similarity and of relatedness. Therefore, a comparison of system performance on the Australian and the two RTÉ datasets provides a unique hindsight into what kind of ratings the Australian respondents assigned. This comparison also gives a vital insight into the inclination towards similarity or relatedness judgments of the system I designed and implemented.

[6]From personal correspondence on the subject with Dr. Lee.

8.3 Threshold Estimation

The binary nature of the task of similarity detection, as it is defined in this thesis (see section 8.1), entails setting two thresholds:

1. **Human Data Threshold**: the respondents in the RTÉ Gold Standard experiment, discussed in chapter 7, were asked to assign a similarity rating for text pairs in the range 0–10 where 0 is identical and 10 is completely different. Conversely, in the Australian news experiment, outlined in section 8.2.3 and discussed in detail in (Lee *et al.* 2004), the rating scale was from 1–5, where 1 was completely different and 5 identical. It was necessary to set a threshold for the human data to convert the range of values to a binary judgment of similarity *versus* dissimilarity. Given the somewhat different test methodology used, the process of threshold estimation was carried out on the Australian and the RTÉ data separately.[7]

2. **Similarity Measure Threshold**: the Jaccard, Dice and Asymmetric similarity measures output numbers in the range 0–1, where 0 means completely dissimilar and 1 means identical. Again, it was necessary to set a threshold, $thresh_{sim}$, such that:

   - $0 \leq thresh_{sim} \leq 1$
   - for a value x derived from a comparison of texts $text_a$ and $text_b$, if $x \geq thresh_{sim}$, then $text_a$ and $text_b$ are similar, if $x < thresh_{sim}$, then $text_a$ and $text_b$ are not similar.

Given the different nature and assumptions of the three similarity measures and also of the seventeen comparison techniques discussed in sec-

[7]See section 8.2.4 for a discussion of differences in the Australian and RTÉ datasets.

tions 6.7.3 and 6.6.2, a unique threshold was sought for each measure—technique combination.

In order to estimate optimum thresholds for both the similarity measures and the human data, a random sample of 30% of the Australian data and 30% of the RTÉ data was taken. For each of the seventeen comparison techniques and for each of the three similarity measures, system performance on the two random datasets was tested at all human data threshold settings within a critical range and at all similarity measure threshold settings. The critical range for the human threshold were 1–2.5 for the Australian data and 6–10 for the RTÉ data. The optimum threshold values were those for which the similarity and dissimilarity detection F scores were highest.[8]

8.3.1 Results of Threshold Estimation Process

The outcome of this threshold estimation process on the subsets of randomly selected data yielded interesting results.

Extreme and balanced threshold values

For both datasets, the system performed best under two very different conditions:

1. **Extreme threshold values**: Where the similarity measure thresholds and, in the case of the Australian data, the human data thresholds were assigned extreme values, the system overall performed well. However, in these cases, results were skewed towards making dissimilarity judgements. This mirrors a distinct skew in the human data towards judging text pairs as dissimilar.

[8]See section 8.4 for an explanation of the performance metrics used.

2. **Balanced threshold values**: The system also performed well overall where both the human and similarity measure thresholds were set to less extreme values. In this case, the system performed equally well on detecting similar and dissimilar text pairs rather than showing a clear bias towards one or the other extreme.

For this reason, system performance was tested for both extreme and balanced threshold values.

Human Data Thresholds

Two thresholds, therefore, were selected for the Australian data but only a single one for the RTÉ data. The thresholds are set out in table 8.1 where the threshold is expressed on the original experiment scale (1-5 for Australia and 10–0 for RTÉ) and scaled to a value between 0 and 1, where 0 is completely different and 1 is identical.

DataSet	Extreme Threshold	Balanced Threshold
RTÉ	—	7.1
RTÉ (scaled)	—	0.35
Australian	1.4	2.2
Australian (scaled)	0.1	0.3

Table 8.1: Threshold values for Human Data

While it may seem inappropriate to assign different thresholds for the two human datasets, as noted in section 8.2.4, the comparability of the two datasets is not a given. In addition to this, the difference in the spread of values in the two sets of human data suggest that respondents in the two experiments were using a very different scale. The Australian data showed a

quite even spread skewed towards dissimilarity. While in the RTÉ data, the skew towards dissimilarity was much more pronounced as respondents barely made use of the 0-5 end of the scale at all. These factors justify the use of different thresholds for the two datasets.

Similarity Measure Thresholds

Each of the seventeen techniques aim to emphasise different aspects of the text representation graph structures. Therefore, it was expected that for each technique a technique-specific similarity threshold would be required. During the course of the threshold estimation process, it also became clear that the three similarity measures have very different ranges for optimum threshold values. For all techniques, the range of values for an optimum threshold for the Jaccard measure was lower than for both the Asymmetric and Dice measures.[9] This is to be expected as, in contrast with the other two measures, the Jaccard measure only counts set intersection once relative to both set difference counts. The optimum thresholds for the Asymmetric measure, which evaluates each set independently of the other, were overall slightly higher than for the Dice measure. Despite the difference in the range of threshold values for the three similarity measures, the optimum values are almost perfectly correlated. Figures 8.1, 8.2 and 8.3 illustrate this phenomenon for the three threshold sets which will be used in the evaluation:

1. **Australian**: optimum threshold set for the seventeen techniques for **balanced** system performance, derived from a random selection of the Australian data

2. **Australian Extreme**: optimum threshold set for the seventeen tech-

[9]See section 6.7.3 for a discussion of these measures.

niques for **extreme** system performance, derived from a random selection of the Australian data

3. **RTÉ**: optimum threshold set for the seventeen techniques for **balanced** system performance, derived from a random selection of RTÉ data

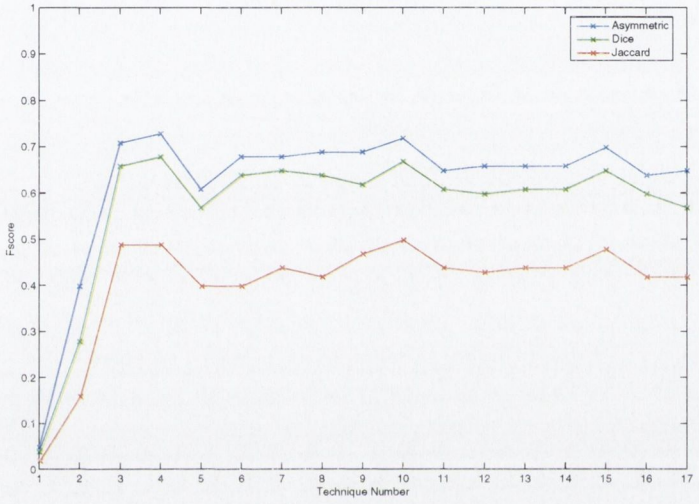


Figure 8.1: Australian: Line Plot of Optimum Thresholds for Jaccard, Dice and Asymmetric measures

8.3.2 Final Evaluation Test Cases

The results in this chapter report system performance separately for each of the three data sets: RTÉ, RTÉ Rel and Australian. The final outcome of the threshold estimation process reported in this section provides three separate test cases for the three datasets based on different combinations of human data and similarity measure threshold types. Each dataset is tested with

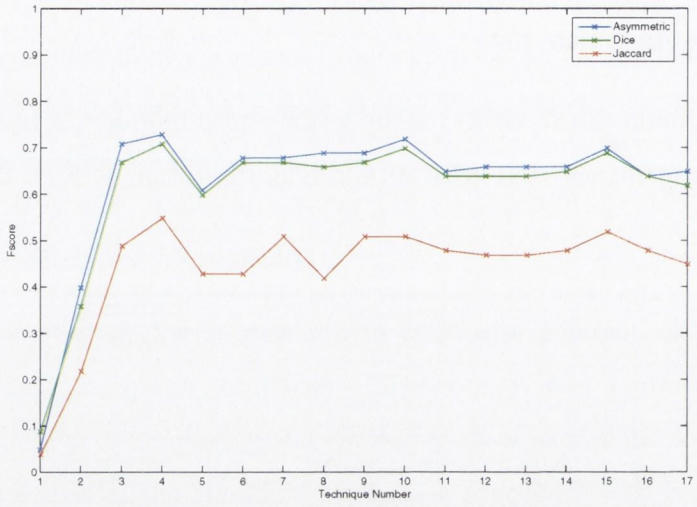


Figure 8.2: Australian Extreme: Line Plot of Optimum Thresholds for Jaccard, Dice and Asymmetric measures

a human data threshold derived from that dataset. That is, for all three threshold test cases (Australian, Australian Extreme and RTÉ), the RTÉ dataset is tested using the human data threshold of 7.1 derived from the RTÉ data, the RTÉ Rel dataset is tested using a threshold of 7.5 derived from the RTÉ Rel dataset and the Australian dataset is tested using a human data threshold derived from the Australian dataset. The threshold combinations for each test case are as follows:

Test Case 1: Australian

- balanced human data thresholds
- balanced similarity measure thresholds, derived from Australian random data subset

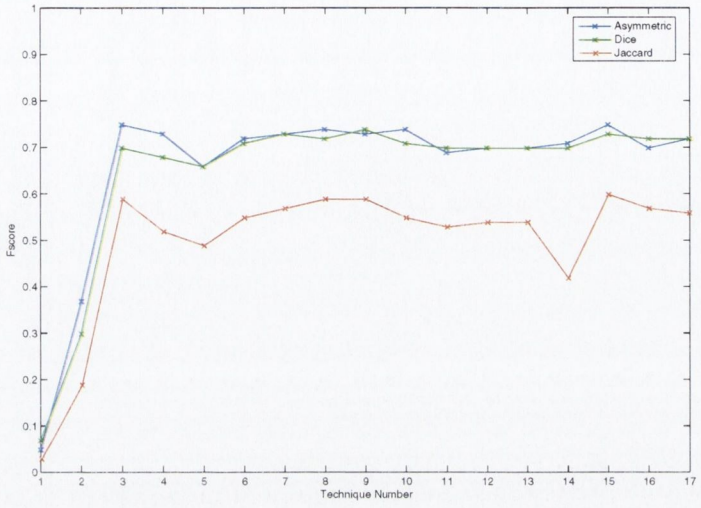


Figure 8.3: RTÉ: Line Plot of Optimum Thresholds for Jaccard, Dice and Asymmetric measures

The Australian human data threshold is the relevant balanced threshold reported in table 8.1.

Test Case 2: Australian Extreme

- extreme human data thresholds
- extreme similarity measure thresholds, derived from Australian random data subset

The Australian human data threshold is the relevant extreme thresholds reported in table 8.1. This combination of thresholds is extremely biased in favour of dissimilarity.

Test Case 3: RTÉ

- balanced human data thresholds
- balanced similarity measure thresholds, derived from RTÉ random data subset

Test Case 2, **Extreme**, is designed to examine how the system performs when biased towards dissimilarity judgments. A contrast of the results for Test Cases 1, **Australian**, and 3, **RTÉ**, is intended to highlight two factors. Firstly, given the difference in dataset size, the assumption is that thresholds based on Australian data would perform better than those based on the much smaller RTÉ dataset. Secondly, an analysis of the performance of the RTÉ Rel dataset on these test cases is intended to investigate whether it correlates better with the RTÉ similarity data or the Australian data. If using the Australian thresholds yield better performance, there would be some basis for suggesting that the Australian data is a measure of relatedness rather than similarity. The next section discusses the performance metrics by which the system has been evaluated for these test cases.

8.4 Performance Metrics

The final results reported below evaluate system performance on three datasets for three separate test cases, three similarity measures and seventeen techniques.[10] This is a total of $3 \times 3 \times 17$ permutations of the system. The aim is to determine which of these 153 permutations is the most successful in identifying similar *and* dissimilar text pairs. Although chapter 6 set out the basic task of the system as one of measuring and determining text *similarity*, in fact, system performance must be judged on this task and the complimentary

[10]Table 8.2 has been provided as a point of reference for the reader giving each technique name and number. See table 6.6 for the features used in weight calculation.

task of determining dissimilarity.

#	**Technique Name**
1	Count
2	Frequency
3	Frequency No Polysemy
4	Factor Out Polysemy
5	InText
6	InText Factor Out Polysemy
7	Paraphrase
8	Repetition
9	All Text Values
10	Arc Type
11	Node Specificity
12	Normalized Node Specificity
13	Arc Type and Node Specificity
14	Arc Type and Normalised Node Specificity
15	Arc Type and All Text Values
16	All
17	All with Normalised Node Specificity

Table 8.2: Seventeen System Techniques

The metric used to determine performance on these tasks is the F Score measure which is a combination of the traditional Information Retrieval evaluation measures of precision and recall.[11] For performance on the similarity

[11]The number of test parameters, 153, makes an analysis using both precision and recall scores, rather than a single F score, unwieldy from the perspective of both the writer and the reader. The precision and recall values, however, were examined and are discussed

detection task precision and recall are defined as follows:

- Precision: the number of correct similarity judgments divided by the number of similarity judgments made by the system
- Recall: the number of correct similarity judgments divided by the number of similarity judgments in the human data

A "correct similarity judgment" is an instance where the system returns that a text pair is similar and that text pair is rated as similar in the human data. Likewise, a "correct dissimilarity judgment" is made when the system returns that a text pair is not similar and that text pair has been rated as dissimilar in the human data. The F Score measure combines these two metrics in order to provide an overall evaluation of system performance. This metric was calculated for both similarity and dissimilarity judgments. The formulae for each metric are set out in equations 8.1 to 8.6. In these equations, $Metric_{sim}$ refers to the metric for similarity judgments and $Metric_{dis}$, the metric for dissimilarity judgments.

$$Precision_{sim} = \frac{\text{Correct Similarity Judgments}}{\text{Total Number of Similarity Judgments}} \tag{8.1}$$

$$Recall_{sim} = \frac{\text{Correct Similarity Judgments}}{\text{Number Of Human Similarity Judgments}} \tag{8.2}$$

$$FScore_{sim} = \frac{2 * Precision_{sim} * Recall_{sim}}{Precision_{sim} + Recall_{sim}} \tag{8.3}$$

$$Precision_{dis} = \frac{\text{Correct Dissimilarity Judgments}}{\text{Total Number of Dissimilarity Judgments}} \tag{8.4}$$

$$Recall_{dis} = \frac{\text{Correct Dissimilarity Judgment}}{\text{Number Of Human Dissimilarity Judgments}} \tag{8.5}$$

$$FScore_{dis} = \frac{2 * Precision_{dis} * Recall_{dis}}{Precision_{dis} + Recall_{dis}} \tag{8.6}$$

where they provide additional insight to the evaluation findings.

In light of the fact that the task of similarity and dissimilarity detection is *not* a standard or common task in the domain of information retrieval or natural language processing in general, it was necessary to derive baseline metrics against which to compare system performance. For many NLP tasks such as ambiguity resolution or document retrieval, other systems provide an indication of what constitutes good or poor performance on that task. The literature provides no such point of comparison for the task in question here. In order to provide some objective measure of goodness, therefore, the performance of this system was evaluated relative to two baseline metrics outlined below:

1. performance which would be expected if judgments were made according to chance;

2. performance expected if judgments were made according to a complete dissimilarity bias.

The chance baseline is the F Scores for similarity and dissimilarity based on the system making judgments on a purely random basis. In this case, the probability of judging a text pair to be similar or dissimilar is 0.5. The number of similar or dissimilar judgments made by the system, then, is half of the total judgments in the dataset, as in equation 8.7. The probability of making a correct similar judgment is the product of the probability of the system making a similar judgment, 0.5, and the probability of the judgment being similar in the human data, $\frac{\#\,Human_{sim}}{\#\,Judgments}$. The number of correct judgments, therefore, is derived according to equation 8.9. The precision and recall values are calculated using these chance estimated values for the number of system similarity or dissimilarity judgments and the number of these judgments that were correct. The formulae used to calculate the chance

precision and recall values are as above (equations 8.1, 8.2, 8.4 and 8.5). Equations 8.10 to 8.13 set out the derivation of how these were calculated from the chance estimated judgment counts for this evaluation. The F Scores are calculated as above (equations 8.3 and 8.6).

$$\#\ System_{sim} = \#\ Judgments \times 0.5 = \frac{\#\ Judgments}{2} \tag{8.7}$$

$$\begin{aligned} \#\ Correct_{sim} &= \frac{\#\ Judgments}{2} \times \frac{\#\ Human_{sim}}{\#\ Judgments} && (8.8) \\ &= \frac{\#\ Human_{sim}}{2} && (8.9) \end{aligned}$$

$$\begin{aligned} Chance\ Precision_{sim} &= \frac{\#\ Human_{sim}}{2} \times \frac{2}{\#\ Judgments} && (8.10) \\ &= \frac{\#\ Human_{sim}}{\#\ Judgments} && (8.11) \end{aligned}$$

$$\begin{aligned} Chance\ Recall_{sim} &= \frac{\#\ Human_{sim}}{2} \times \frac{1}{\#\ Human_{sim}} && (8.12) \\ &= 0.5 && (8.13) \end{aligned}$$

The dissimilarity bias baseline is based on the system assigning a dissimilar judgment for *all* text pairs. In this case, $F\ Score_{sim}$ is always zero and $Precision_{dis}$, $Recall_{dis}$ and $F\ Score_{dis}$ are calculated according to the following formulae:

$$\begin{aligned} Precision_{dis} &= \frac{\#\ Human_{dis}}{\#\ Judgments} && (8.14) \\ Recall_{dis} &= \frac{\#\ Human_{dis}}{\#\ Human_{dis}} = 1 && (8.15) \\ F\ Score_{dis} &= \frac{2 \times Precision_{dis} \times Recall_{dis}}{Precision_{dis} + Recall_{dis}} && (8.16) \\ &= \frac{2 \times Precision_{dis}}{1 + Precision_{dis}} && (8.17) \end{aligned}$$

System performance, then, represented by the similarity and dissimilarity F Score values calculated on the 153 system permutations was evaluated relative to the F Scores for the chance baseline and the F Scores for the dissimilar-bias baseline. The following section sets out how these metrics are presented in a series of graphs to illustrate system performance.

8.4.1 Graphic representation of results

For each dataset, there are three graph sets. Each set contains nine graphs, one for each Test Case and Similarity Measure combination. Each row in the set of graphs presents a Test Case and each column a similarity measure.[12] In the first graph set (figures 8.4, 8.7 and 8.10), each graph plots the similarity and dissimilarity F Scores (fs and fd, respectively), the chance F scores ($chancefs$ and $chancefd$) and the dissimilar bias F Scores ($Alldisfs$ and $Alldisfd$) for the seventeen system techniques. These graphs show how the system performs relative to the two baselines. The second graph set (figures 8.5, 8.8 and 8.11), again, containing nine graphs, plots the binary result of filtering the results for system performance on similarity *and* dissimilarity greater than chance with a tolerance of 0. In this graph set, any technique that performs better than the chance F Scores has been assigned a value of one and those that do not have been assigned a value of zero. This graph set therefore illustrates those technique, Test Case, similarity measure combinations that perform better than chance. The final graph set (figures 8.6, 8.9 and 8.12) again plots the result data filtered for performance greater than chance. In this set, however, the actual F scores are plotted, not a binary result. These graph sets, then, illustrate, for those technique, Test Case and

[12]Rows and columns mentioned here are based on the assumption of turning the page sideways into a reading position.

similarity measure combinations that perform better than chance, *how much* better than chance they perform. The following section describes the results, illustrated using these graph sets, and discusses their implications for an evaluation of system performance.

8.5 Similarity Detection Results

This section discusses how the system I designed and implemented performs on the three datasets: RTÉ, Australian and RTÉ Rel. As stated in section 8.4, there are seventeen techniques by three similarity measure by three threshold combinations, giving a total of 153 system permutations in total, to be evaluated. The aim here is to determine which permutations perform better than chance consistently across the datasets. The dissimilarity bias baseline is primarily included in this presentation to give an indication of the degree of skewness of the human data towards dissimilarity as this baseline is always significantly higher than chance. The results for each of the datasets are first presented individually in sections 8.5.1, 8.5.2 and 8.5.3. Section 8.5.4 synthesises the results for all three datasets, highlighting the techniques, similarity measures and threshold test cases which proved most and least successful overall. Some implications of these results for a definition of text similarity, for gold standard experiment design and for applications based on the notion of similarity are discussed in section 8.5.5. In the discussion and graphs that follow, techniques are often referred to by number. The reader, therefore, should use table 8.2 as a point of reference for technique name and number.

8.5.1 RTÉ Gold Standard Evaluation Results

Figures 8.4, 8.5 and 8.6 present the similarity and dissimilarity F Scores and base line metrics for system performance on the 192 text pair RTÉ dataset. A first point to note in figure 8.4 is that, for all parameter combinations, the system is much more accurate in detecting dissimilarity than similarity with the maximum dissimilarity F score at 0.915 and the maximum similarity score at 0.381.[13] Furthermore, the skew in the gold standard data towards dissimilarity is apparent in the low chance baseline for similarity contrasted with the high chance baseline for dissimilarity. Therefore, both the system and the human subjects are biased towards dissimilarity. The following sections set out points to note regarding test cases, similarity measures and techniques.

Test Cases

While the similarity detection scores are relatively constant at slightly higher than chance across the nine sub-graphs in figure 8.4, the dissimilarity F scores show a marked improvement from Test Case 1, Australian, (row 1) to Test Case 3, RTÉ, (row 3). The RTÉ dataset scores higher on Test Case 3 which has been derived from RTÉ data than on Test Case 1 derived from the Australian dataset. This is in line with expectations that datasets would perform better on the Test Cases which are derived from them. Indeed, the scores for Test Case 1 are generally lower than the chance baseline. The Extreme Test Case, Test Case 2 (row 2), however, which is also derived from Australian data, while still lower than Test Case 3, does perform better than chance for two similarity measures. This Test Case is designed to test how the system performs using thresholds that are extremely biased towards dissimilarity.

[13]Dissimilarity results are presented in red and similarity in blue throughout.

The RTÉ gold standard data shows this strong bias and this is reflected in the results for this Extreme Test Case. A glance at figure 8.5 emphasises the clear distinction between performance on the three Test Cases, with Test Cases 2 and 3 showing the greatest number of parameter combinations where similarity and dissimilarity scores are better than chance. Figure 8.6 confirms that for the RTÉ dataset, the RTÉ Test Case, Test Case 3, provides the best combination of parameters with dissimilarity scores, in particular, significantly higher than chance for many techniques. Test Case 2, although less marked, also shows a large improvement on the chance baseline.

Similarity Measures

The major point to note with regard to the different similarity measures is that the Asymmetric measure (column 3) is the least successful. Figure 8.6 illustrates that for Test Cases 1 and 2, no technique performs better than chance with this similarity measure. Even when combined with the RTÉ derived thresholds in Test Case 3 (row 3, column 3) the improvement on the chance baseline is much lower than for the Jaccard and Dice measures. There does not appear to be any significant difference between the Jaccard and Dice measures.

Techniques

A group of techniques that appear to perform quite badly in this dataset are techniques 6–9, based on repetition and paraphrase weighting algorithms. This is a trend which is repeated in the other datasets and is addressed in section 8.5.4.

A second point to note is the large drop in F score value for technique 14 in Test Case 3. This is can be traced to an unusual threshold assignment,

where the optimum performance threshold for the Jaccard measure is not in line with the other measures or Test Cases (see figure 8.3). In fact, following the threshold estimation process, two optima were established, the second option being one which was more closely correlated with other measures. This is a case of over-fitting the thresholds to the data and illustrates, firstly, the sensitivity of the system to threshold values and, secondly, the relative accuracy of the threshold estimation process.

8.5.2 Australian Gold Standard Evaluation Results

Figures 8.7, 8.8 and 8.9 present the similarity and dissimilarity F Scores and baselines for the Australian dataset of 1,225 text pairs. The following trends, outlined for the RTÉ dataset, recur in these results:

1. Dissimilarity detection consistently outperforms similarity detection.
2. Similarity detection performance is more or less constant, achieving a score slightly above chance.
3. The Asymmetric similarity measure performs the worst of the three measures, outperforming chance only marginally or not at all.
4. Techniques 6–9 show a consistent drop in performance relative to other techniques.
5. The maximum dissimilarity F score is 0.885.

The following sections outline features of the results which are specific to the Australian dataset.

Test Cases

In direct contrast to the RTÉ results and in line with expectations, the results for the Australian dataset, illustrated in Figure 8.7, are better for Test Cases 1 and 2, based on thresholds derived from the Australian data than for Test Case 3, based on RTÉ—derived thresholds. In all RTÉ Test Case (Test Case 3) result data, only four techniques outperform chance, and then only marginally. Interestingly, dissimilarity detection F score values remain constant marginally higher than chance but the system fails on the task of similarity detection using RTÉ-derived threshold values.

The contrast between the results for Test Cases 1 and 2, the Australian and Extreme Test Cases respectively, is also striking. The chance F score values for the two datasets are different because each Test Cases uses a different threshold not only for the similarity measure but also for the gold standard data. The more extreme human data threshold value in Test Case 2 pushes the chance F scores to more extreme values with a clear bias towards dissimilarity.[14] It would appear from figure 8.8 that the two Test Cases work equally well, with a similar distribution of techniques outperforming chance. This trend, however, is clarified in figure 8.9 where it is clear that, for the Jaccard and Dice similarity measures, the Extreme Test Case, 2, outperforms chance by a far wider margin on dissimilarity detection. Similarity detection F Score values for Test Case 1, on the other hand, are much higher than for Test Case 2, although only marginally outperforming chance. The two Test Cases represent a choice for potential users between moderate performance on both similarity and dissimilarity detection tasks or excellent performance on dissimilarity and mediocre, but better than chance, performance on similarity detection.

[14]The greater the bias towards dissimilarity, the higher the chance dissimilarity F Score.

Techniques

As noted above, the trend in the RTÉ data towards low performance on repetition-based techniques 6–9 would seem to be repeated in the Australian data. However, it is worth noting that in Test Case 1, Australian, using the Jaccard similarity measure (row 1, column 1 in Figure 8.7), this drop in performance for dissimilarity detection is compensated for by a corresponding improvement in performance in similarity detection, achieving the highest similarity F score value of 0.687. This would suggest a possible hypothesis that the text-based measures strongly bias in favour of similarity. Section 8.5.4 returns to this issue in more detail.

A secondary point to note, again in direct contrast with the RTÉ data, is the lone satisfactory performance of technique 14 in Test Case 3 for the Jaccard Similarity measure (row 3, column 1 in Figure 8.8). It would appear that this threshold value, which is over-fitted to the RTÉ data, is appropriate to the Australian dataset. A comparison of figures 8.1 and 8.3 illustrates that the range of threshold values for Test Cases 1 and 3, Australian and RTÉ respectively, differ by approximately 0.1, with Australian values consistently lower. The technique in question, number 14, alone falls within the range of the Australian threshold values. This observation serves to re-emphasise the sensitivity of the system to the process of threshold determination.

8.5.3 RTÉ Rel Gold Standard Evaluation Results

Figure 8.10, 8.11 and 8.12 present the system and baseline F Scores for the RTÉ Rel dataset.[15] This dataset was included in this evaluation process to test whether the Australian gold standard data in fact modelled text

[15]The RTÉ Rel dataset was tested using a single human data threshold of 7.5 that was derived in a small scale threshold estimation procedure.

relatedness rather than text similarity. The RTÉ gold standard data provides a unique insight into this question as the experiment aimed to separate out these two factors. One might expect, if the Australian data did represent relatedness rather than similarity, that the RTÉ Rel data would perform better in Test Cases 1 and 2, derived from the Australian dataset, than the basic RTÉ dataset. This proves to be the case to a limited extent. A comparison of figures 8.11 and 8.5 illustrate that, for the Australian Test Case, Test Case 1, (row 1 in both figures), five extra techniques perform better than chance on the RTÉRel data. However, an analysis of figures 8.12 and 8.6 show that this improvement is only marginal relative to the RTÉ data, with the techniques in the RTÉRel graphs only marginally better than chance. Likewise, the RTÉRel data performs marginally worse on the RTÉ Test Case 3 thresholds with seven fewer techniques performing better than chance.

This analysis, however, omits a key point of contrast between the RTÉ Rel and the RTÉ data set. The similarity F Score value for the RTÉ Rel dataset for the two Australian Test Cases is in the range 0.4–0.5 while the similarity F Score values for the RTÉ data is in the range 0.25–0.35. This is an improvement of 0.15 on average for the RTÉ Rel dataset on the similarity detection score of the original RTÉ data. It would appear that similarity in the Australian data is more closely correlated with relatedness in the RTÉ data than with similarity. This would strongly support the hypothesis that "similarity" in the Australian dataset, which was defined as similarity of topic, is, in fact, closely linked to what was defined as relatedness in the RTÉ experiment in collecting data from human participants.[16]

[16]See section 7.3.1 for a discussion of the motivations for including two scales in the RTÉ experiment and section 8.2.4 for a discussion of the differences between the Australian and

In other aspects, the results for this dataset display similar characteristics to the RTÉ dataset results with techniques 6–9 and the Asymmetric similarity measure showing poor results and Test Cases 2 and 3 giving best performance.

RTÉ task presentation.

Figure 8.4: RTÉ Data. Unfiltered.

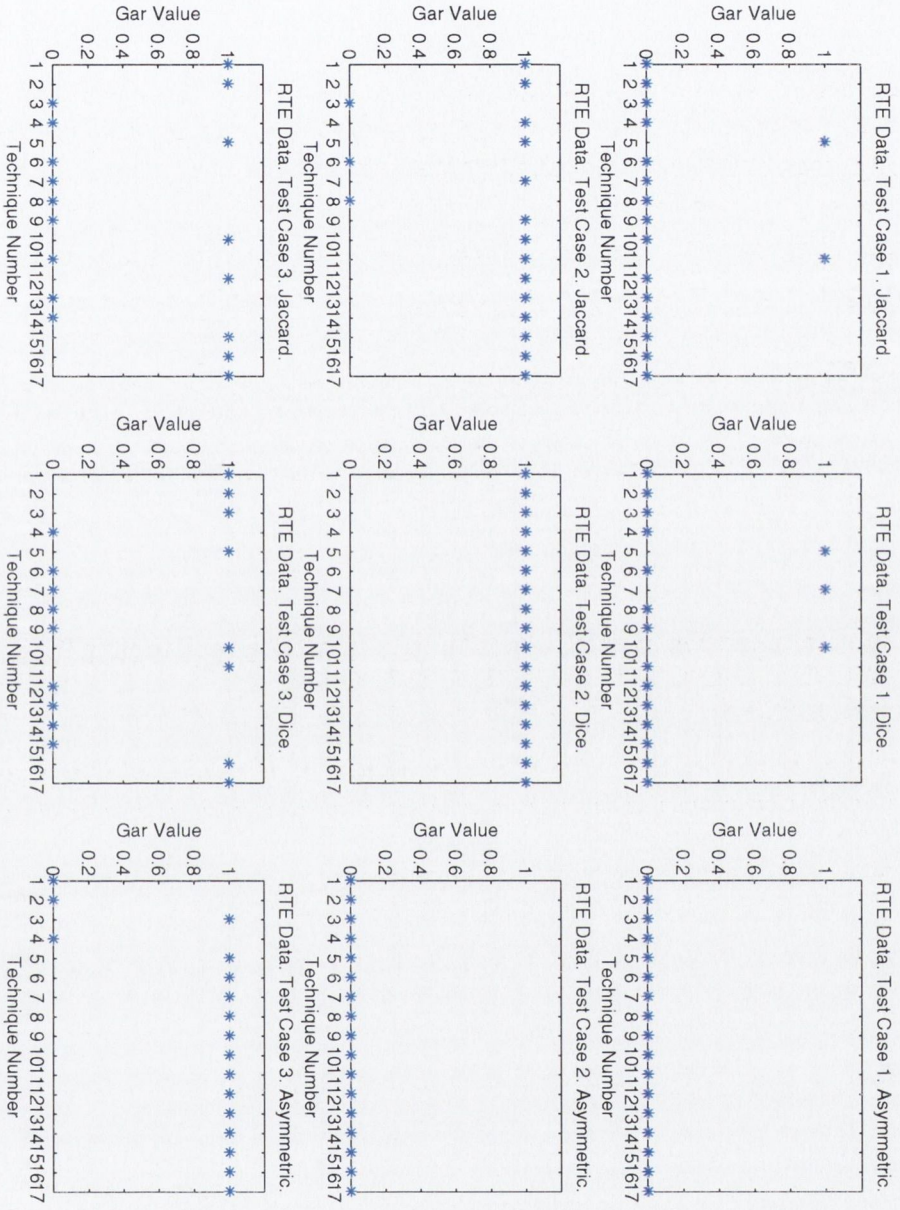


Figure 8.5: RTÉ Data. Binary result when tolerance is 0.

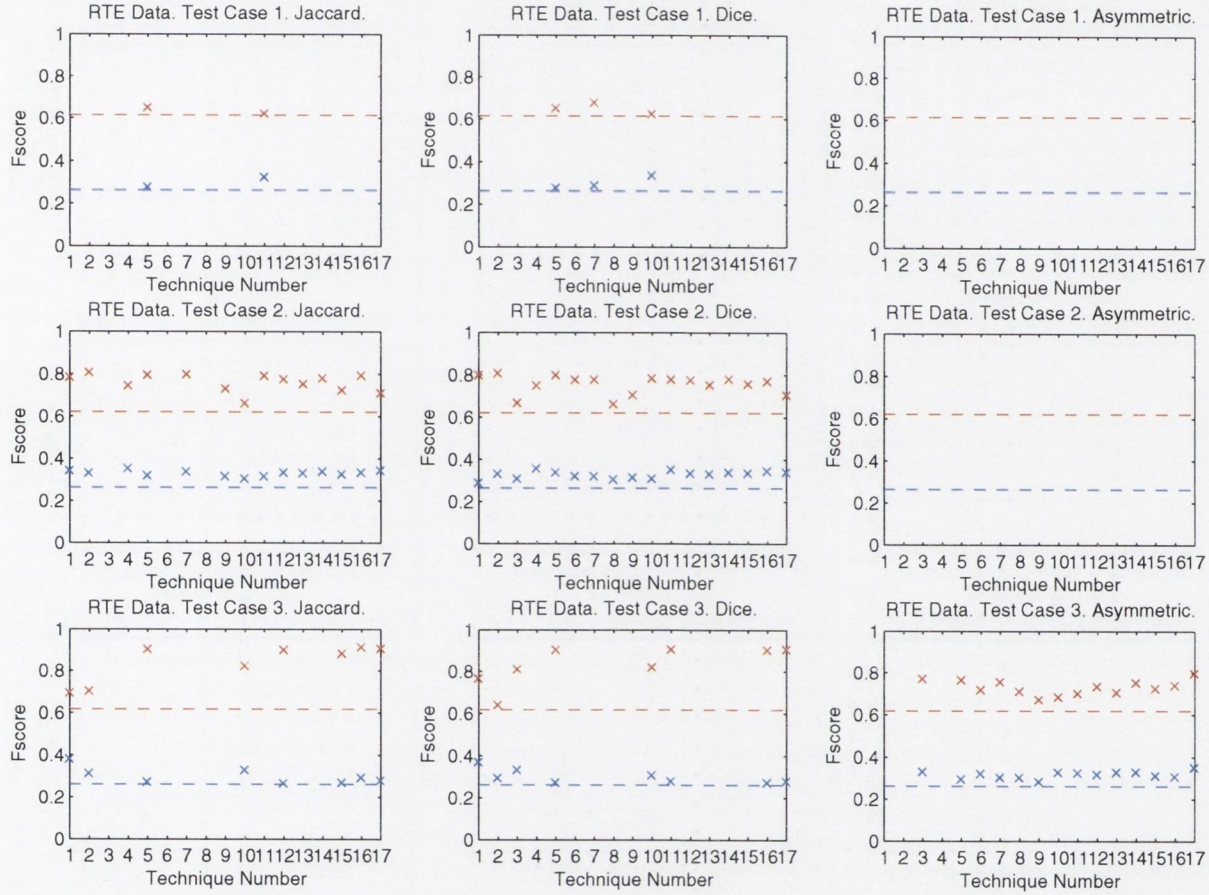


Figure 8.6: RTÉ Data. Filtered when tolerance is 0.

Figure 8.7: Australian Data. Unfiltered.

Figure 8.8: Australian Data. Binary result when tolerance is 0.

Figure 8.9: Australian Data. Filtered when tolerance is 0.

Figure 8.10: RTÉ Rel Data. Unfiltered.

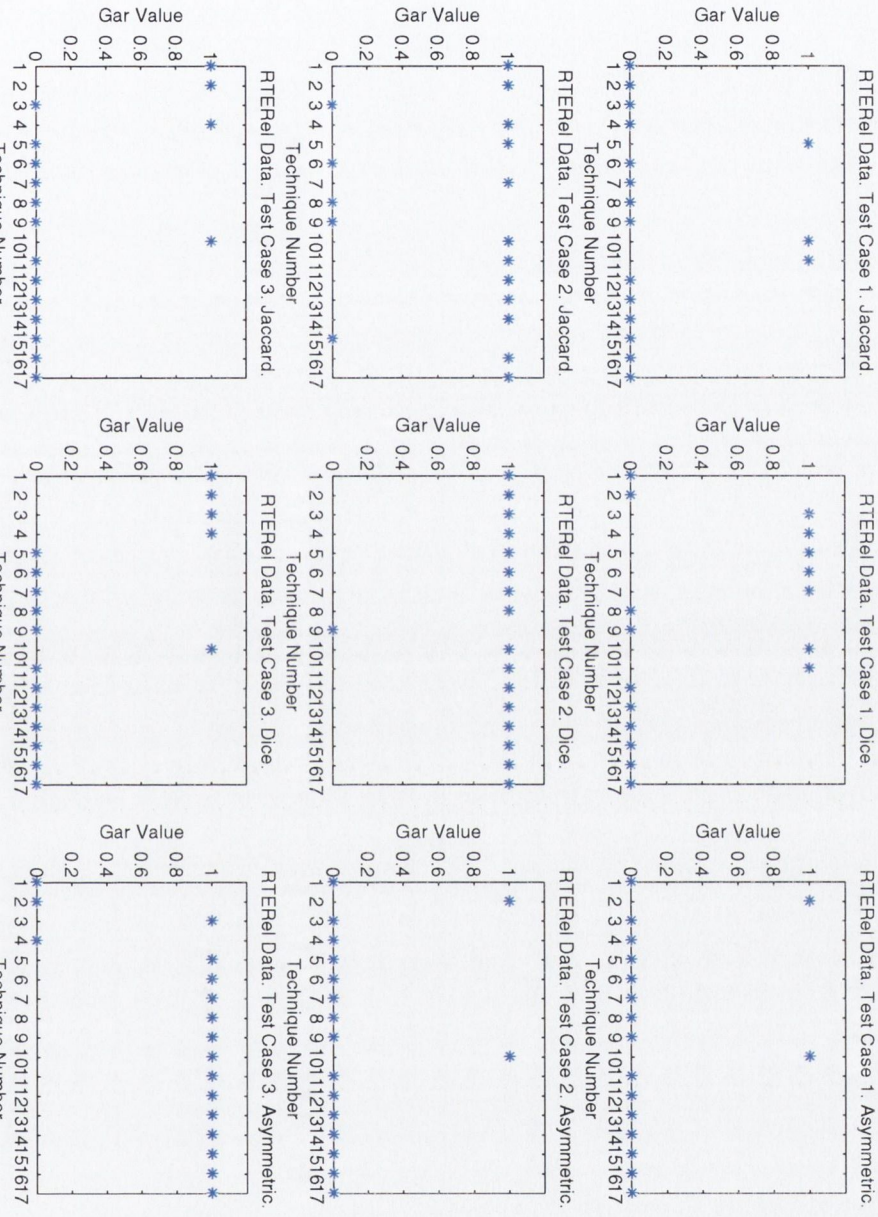


Figure 8.11: RTÉ Rel Data. Binary result when tolerance is 0.

Figure 8.12: RTÉ Rel Data. Filtered when tolerance is 0.

8.5.4 Synthesis of Results

This section is intended to synthesize the results for the RTÉ, Australian and RTÉ Rel datasets to establish which technique, similarity measure and thresholds seem to perform consistently best across the three datasets.

Thresholds and Similarity Measures

The individual dataset analyses of the previous three sections would seem to suggest two hypotheses:

1. the Extreme Test Case, Test Case 2, appears to have consistently high F Score values relative to chance.

2. The Asymmetric similarity measure consistently under-performs.

Table 8.3 summarises the three binary result figures 8.5, 8.8 and 8.11. The counts in each cell refer to the number of techniques for a particular dataset, similarity measure and Test Case whose similarity and dissimilarity F Scores were greater than chance. A clear pattern emerges which supports the hypotheses posited above. The Extreme Test Case appears to most successfully out-perform chance. It also proved to be robust on extension to datasets on which it was not trained (RTÉ and RTÉRel). The Asymmetric similarity measure under-performs consistently relative to the Jaccard and Dice measures, with the Dice measure giving marginally better results than the Jaccard similarity measure. In a system which is biased in favour of dissimilarity, the Dice similarity measure, which counts set intersection weight, i.e. similarity, twice, constitutes a necessary counter-balance in favour of similarity judgments which is lacking in the Jaccard measure. Indeed, the subtotal for the Dice—Extreme combination is forty-six successful techniques out of a possible total of fifty-one, providing further evidence to support adopting

the Extreme thresholds with the Dice similarity measure as the optimum threshold — similarity measure combination.

Test Case	Dataset	Jaccard	Dice	Asymmetric	Total
1 Australian					
	Australian	8	8	2	18
	RTÉ	2	3	0	5
	RTÉ Rel	3	7	2	12
	SubTotal	13	18	4	**35**
2 Extreme					
	Australian	12	13	11	36
	RTÉ	14	17	0	31
	RTÉ Rel	12	16	2	30
	SubTotal	38	46	13	**97**
3 RTÉ					
	Australian	1	0	3	4
	RTÉ	7	8	14	29
	RTÉ Rel	4	5	14	23
	SubTotal	12	13	31	**56**
	Total	**63**	**77**	**48**	

Table 8.3: *Performance Better Than Chance* Count for Test Case and Similarity Measure Combinations

Techniques

Table 8.4 sets out for each technique the number of instances that technique generated a similarity and dissimilarity F Score greater than the values pre-

dictable according to chance. The second count column gives the number of instances which use the Dice and Jaccard similarity measures that were greater than chance, as the Asymmetric measure has been discounted. A

Technique	# Successes	# minus Asymmetric
1	15	12
2	15	12
3	13	9
4	13	10
5	13	10
6	6	3
7	10	8
8	7	3
9	9	5
10	**19**	**13**
11	12	9
12	13	9
13	11	8
14	11	7
15	10	6
16	12	8
17	12	8
Theoretical Max.	**27**	**18**

Table 8.4: Successful Results per Technique

single most successful parameter emerges from this combined analysis of the three data sets: **Technique 10**. This technique assigns weights based on link

type in the text representation graph.[17] and outperforms chance in nineteen out of twenty-seven cases. The success of this technique illustrates the importance of the graph structure in evaluating graph intersection and difference for text comparison.

A second feature of the text representation graph structure which does not appear to have had its expected impact on the comparison process, however, is node specificity, techniques 11 and 12. These performed slightly worse than the basic node and arc frequency counts, techniques 1 and 2. The similarity measure calculation is based on the weights assigned to the intersection and difference graphs. In order to investigate the effect of individual techniques, an analysis of the intersection and difference weight values for each text pair in the RTÉ corpus was carried out. It emerged that increases or decreases in the weights of the intersection and difference graphs were not correlated. That is, an increase or decrease in the weight of the intersection graph from one technique to another does not entail a systematic corresponding increase or decrease in the weight of the difference graphs. This is the case for all techniques, *except* the Node Specificity techniques 11 and 12. The set weights for these techniques are highly correlated with those for all others. The node specificity weighting algorithm appears to have the *same* increasing effect on all graphs or sub–graphs. The expectation was that intersection graphs might contain nodes which were less specific and this would have a significant impact on the overall node specificity weight for intersection graphs. Conversely, difference graphs were expected to contain nodes which were more specific and this would have a corresponding effect on the node specificity weight. This, however, does not appear to be the case which would suggest that the intersection and difference sub-graphs have a similar

[17]See section 6.6.1 for details of the link type weighting algorithm.

distribution of node types.

A second technique type proved less successful than expected in the task of similarity detection: techniques 6–9. Previous sections noted that these techniques overall did not perform well. Table 8.4 localises this poor performance to techniques 6, 8 and 9 in particular and to a lesser extent techniques 7 and 15. All the afore-mentioned techniques are relevant for graph nodes representing terms that appear in the source texts. Technique 6 aims to factor out the multiplicative effects of polysemy for words that appear in the text. Techniques 7 and 8 assign greater weights for paraphrase and repetition respectively. Technique 9 is a combination of techniques 8 and 9. Finally, technique 15 combines technique 9 with the link type weighting algorithm of technique 10. Section 2.6 stressed the importance of repetition and paraphrase as a lexical cohesive device and section 2.6.3 noted how repetition is the most powerful factor in determining lexical chain membership. It is, therefore, surprising that text-based cohesive devices should have such a negative effect on the performance of this system for text comparison. Section 8.5.2 posited the hypothesis that text-based cohesive devices bias the system towards similarity judgments contrary to the dissimilarity bias which is evident in human judgments and in other system techniques. In fact, this hypothesis is borne out by a more in-depth analysis of the performance metrics for these techniques. The Precision value for dissimilarity detection using these techniques stays more or less constant relative to other more satisfactory techniques but the Recall value drops dramatically. This is due to an upsurge in the number of similarity judgments made by the system and a corresponding drop in dissimilarity judgments, thereby, adversely effecting the correct dissimilarity to real dissimilarity judgment ratio, $Recall_{dis}$. This would suggest that the text-based cohesive device weighting algorithms could

be modified to take account of an apparent over—weighting of repeated terms in the intersection graph and a corresponding under—weighting in difference graphs.

A consistent pattern emerges from all datasets. Best performance is attained using the Extreme threshold Test Case 2 combined with the Dice similarity measure and using system Technique 10 based on graph link type weights. The Asymmetric similarity measure proved least useful. The RTÉ and balanced Australian thresholds proved not to be extendible to datasets other than those on which they were trained. Finally, the relatively poor performance of the node specificity and text-based cohesive device techniques suggests that the distribution of nodes in the intersection sub-graph is not restricted to mainly generic concepts *derived* from text rather than actually present in the text. Instead, it covers generic and specific concepts and nodes derived from text but also those present and indeed repeated in the source texts.

8.5.5 Discussion of text similarity

Section 8.5.4 compares system performance for the three gold standard datasets synthesising the results for all similarity measures, techniques and Test Cases. This section raises some questions regarding the notion of text similarity that have emerged during the system performance evaluation process. This issue mainly arises from the comparison of results for the similarity RTÉ dataset and the relatedness RTÉ Rel dataset in section 8.5.3. It appears that the notion of similarity derived from the Australian data has more in common with the notion of relatedness in the RTÉ Rel data than that of similarity in the RTÉ data. The three datasets may in fact represent three slightly different concepts:

1. **Text similarity in the RTÉ dataset**: a concept which is very difficult to define because text is a complex structure and similarities can be drawn between such complex structures on many levels. This difficulty is reflected in the low inter-rater agreement scores in the RTÉ gold standard experiment and in the feedback from respondents on the difficulty of the task of assigning similarity ratings.

2. **Topic similarity in the Australian dataset**: this concept is clearly defined. This clarity is reflected in the high inter-rater agreement scores for the Australian gold standard task and the lack of negative feedback. This reduces text similarity to a one-dimensional comparison of topic area and ignores the complexity inherent in text.

3. **Text relatedness in the RTÉ Rel dataset**: this concept is related to text similarity[18] but appears to be more closely related to topic similarity, given the similarity detection scores for the RTÉ Rel dataset on the Australian threshold settings.

This difficulty in defining a clear notion of what it means for two texts to be similar has implications for both experiments eliciting human judgments such as that described in chapter 7 and for applications which are based on the notion of similarity, such as that described in chapter 6 of this thesis. For example, an application that deals with semantic orientation of text judges similarity not in terms of topic but in terms of writer bias. Conversely, a text categorization application may be concerned primarily with similarity of topic and not bias, style or any other aspects of text. In an experiment to derive a gold standard to test such systems, the specific notion of similarity must be made explicit in order to elicit appropriate gold standard judgments

[18]See section 7.3.1 regarding respondents comments on similarity and relatedness.

from respondents. This discussion does not aim to invalidate any definition of text similarity but rather to highlight the necessity of specifying that definition in empirical investigation. The results from the three datasets used in this evaluation illustrate how subtle differences in this definition can produce significant and perhaps unexpected effects.

8.6 Difference Identification Results

The Storylines component to extract the sentences which contain the most differential information content between two texts was evaluated in a small-scale experiment. A pilot gold standard of the two main differential sentences for eight text pairs was derived from an experiment to elicit differential sentence judgments from human subjects. Section 8.6.1 describes the gold standard experiment conducted. Section 8.6.2 discusses the performance of the Storylines system on this derived pilot gold standard.

8.6.1 Google News Difference Identification Experiment

Materials

The test corpus evaluated by the human judges was a collection of eight text pairs taken from the Google News website.[19] This site groups articles from on-line newspapers around the world by similar or related content. The texts selected cover seven different news topics and are on average 330 words and 13 sentences long.[20]

[19] I developed a tool to trawl the Google News website, download groups of related articles and strip them of all html mark-up. The corpus collection process was, therefore, automated.

[20] The texts are set out in appendix A.4.

Subjects

The subjects were 4 adults, 2 male and 2 female, all right-handed, with a mean age of 32 years. They were solicited from among friends and family of the researcher.

Procedure

The subjects were asked to read each of the eight pairs of texts. Each text pair was printed on a separate page and all sentences were numbered, as in Appendix A.4. For each pair, they were required to mark the sentences in the two texts that presented information not provided by the other text. It was stressed that the aim was to identify differences in *content*, not surface vocabulary or structural differences. Subjects were advised to mark only the three or four most different sentences in the texts. The experiment was carried out under supervised conditions and took approximately 60 minutes.

Basic Results

For all texts, there are at least two sentences for which there was unanimous or near unanimous (three out of four) agreement on the most different sentences.[21] These sentences constitute the "gold standard" against which to test the system differential sentence extraction algorithm.

8.6.2 Google News Evaluation Results

The Storylines differential sentence component was tested for the eight story pairs of the experiment set out above. The gold standard consists of two sentences for each text in a text pair giving a total of 32 sentences against

[21] The identified differential sentences are marked in bold in the texts in Appendix A.4.

which to match system results. The results for the seventeen techniques are set out in table 8.5. A key point to note in this evaluation is that results for differential sentence selection according to chance are much lower than 50%. The input texts are on average 13 sentences long. Therefore, if Storylines was selecting output sentences purely at random from the text, the likelihood of selecting two sentences from the gold standard is the sum of the likelihood of selecting first one sentence from the thirteen possible sentence and then the second one from the remaining twelve possible sentences, as in equation 8.18.[22]

$$\frac{2}{13} + \frac{2}{12} = 0.154 + 0.166 = 0.32 \tag{8.18}$$

In the light of a baseline performance of approximately 32%, the results in table 8.5 all show a marked improvement on chance for the eight text pairs. This is quite encouraging and would suggest that this component should be further developed and a larger-scale evaluation of its performance carried out.

A second point to note is that the results would suggest that the task of differential information content extraction is substantially different from the similarity estimation task. The technique, number 10, which performed best for similarity detection performs slightly worse than average on the difference identification task. On the other hand, the basic node presence technique, technique 1, whose performance on similarity estimation was not noteworthy, outperformed all other techniques by a significant margin. It would appear, therefore, that technique performance on the similarity and difference tasks

[22] I stress that this chance baseline is approximate. The texts vary in length from eight to nineteen lines. Therefore, a chance baseline should be calculated for each story and results compared on a text by text basis. For this pilot study, however, the approximate chance baseline is suggested.

Technique	# correct	% correct
Theoretical Max.	32	100%
1	22	68.75%
2	20	62.5%
3	13	40.625%
4	12	37.5%
5	16	50%
6	16	50%
7	17	53.125%
8	16	50%
9	16	50%
10	13	40.625%
11	16	50%
12	15	46.875%
13	14	43.74%
14	14	43.74%
15	14	43.74%
16	16	50%
17	16	50%
Total	266	48.89%
Average	15.65	48.89%
Chance	174	32%

Table 8.5: Differential Sentences Results for 17 System Techniques

is not correlated.

A second analysis was carried out of the results from the perspective of the text pairs involved. The results, set out in table 8.6, suggest a somewhat different interpretation. It is clear from this analysis that the apparent overall satisfactory performance of the seventeen techniques is largely due to the success of the Storylines system on text pairs 5 and 6. These two text pairs regard, respectively, the death of French philosopher, Jacques Derrida, and bombings in the Sinai peninsula in Egypt. In contrast with other text pairs, both the Derrida and Sinai peninsula text pairs present very clear changes in content mid-way through the texts. The Derrida texts both present some salient details of the life of the philosopher. Subsequently, one of the texts goes on to discuss the philosopher's life-long ambition to be a footballer, the other outlines aspects of his theories. Likewise, for the Sinai text pair, both provide an initial account of the aftermath of the bombings. Following this one text expands on hear–say evidence of events before the bombings, the other discusses the Israeli cabinet's response to the events. The other text pairs in the evaluation do not present such clear cut distinctions in content. This would suggest that the difference algorithm is not sensitive to fine-grained content differences but detects rather shifts in overall content focus. This characteristic may be a function of the restricted two sentence output provided by the system where it is unable to suggest fine-grained distinctions, particularly within sentences with broadly similar content.

8.6.3 Google News Experiment Discussion

The results of the pilot evaluation of the Storylines differential sentence content component indicate that the component is successful in detecting large shifts in content but less successful on more fine-grained distinctions. In

Technique	# correct	% correct
Theoretical Max.	**68**	**100%**
1	25	36.76%
2	16	23.5%
3	29	42.65%
4	26	38.24%
5	63	92.64%
6	52	76.47%
7	27	39.71%
8	29	42.65%
Total	266	48.89%
Average	33.25	48.89%

Table 8.6: Differential Sentences Results for 8 Story Pairs

the human experiment, respondents often noted more than two sentences for each text which contained differential information. Some of the distinctions made were quite fine-grained, others less so. The high number of differences noted per text pair would suggest that the articles grouped together under one topic on the google news website in fact contain *a lot* of different information. The Storylines system, however, is restricted to suggesting only the two most differential sentences although there may be several close candidate sentences for selection. These observations suggest two new directions for research on this topic:

- The first new direction relates to experiment design. It proved difficult to restrict the gold standard sentences to two sentences per text against which to compare system performance. An alternative human experiment design could provide a more accurate evaluation of the performance of the Storylines system as it is implemented currently. The experiment could be proposed as an explicit output evaluation experiment where respondents are presented the output of the Storylines system and asked to indicate whether these sentences constitute the main differential content between the two source texts. This experimental approach would also significantly reduce the difficulty of the human evaluation task as the task of evaluation is more focused than the task of sentence selection which comprises an element of both evaluation and discernment.

- A second new direction regards the output of the Storylines system. The point has been made several times that the output of the differential sentence component of Storylines, although helpful for a potential browser of documents, is quite restrictive. This is borne out by empirical evidence in this experiment where it would appear that two

sentences is often either not sufficient or too coarse-grained to indicate the main differences between two texts. The text representation model described in chapter 5 provides rich semantic and structural information which should be exploited to provide more interpretable and informative salient difference output. Section 9.2.5 outlines some possible directions for future development of the output for this Storyline component.

8.7 Discussion

Storylines, reported in chapters 5 and 6, was developed as an implementation of my intuitions regarding possible factors in determining similarity and dissimilarity of texts using WordNet as an added information source. This chapter provides a thorough evaluation of the similarity estimation component of the Storylines system. Section 8.2 sets out the three datasets used for evaluation and highlights some of the key differences between them which motivate the inclusion of all three datasets. The system evaluation is based on the binary task of detecting similar versus completely dissimilar texts. Section 8.3, therefore, describes how thresholds were estimated for the Storylines similarity measure output and the datasets mean rating for use in this binary task. Section 8.4 defines the metrics used to evaluate system performance. Section 8.5, then, provides an in-depth analysis of system performance on the three Test Cases, seventeen Storylines techniques and three similarity measures in the evaluation. This chapter also reports a pilot evaluation of the differential sentence extraction component of the Storylines system in section 8.6.

The chapter has raised some interesting issues regarding the notions of

similarity and relatedness of texts. The comparative evaluation of the RTÉ, Australian and RTÉ Rel data sets in section 8.5.4 would suggest that subtle differences in the definition of the similarity judgment task in psycholinguistic experiments can result in clear differences in the judgments elicited. Further, the differences apparent in the three datasets would suggest that similarity as topic similarity in texts is more closely related to text relatedness than to a general notion of text similarity. This contrast between one-dimensional and multi-dimensional definitions of similarity has repercussions for both experiment and application design. The empirical findings reported in this chapter have served to highlight the importance of providing a fine-grained analysis of what we mean by text similarity. The chapter that follows presents the conclusions and contributions of the research reported in this thesis and presents avenues for future research that lead on from this empirical and computational work.

Chapter 9

Future Work and Conclusions

9.1 Introduction

Chapter 1 of this thesis stated that the guiding principle of this research has been how to represent text meaning, from both a theoretical and a computational perspective. Chapter 2 set out the theoretical grounding of this thesis in current theories and applications of text models in the literature with a particular emphasis on models of lexical cohesion which provide the central framework for my research. Chapter 3 described the materials, the WordNet knowledge base and news text corpora, used in the computational and empirical aspects of this research. The use of a hand-coded knowledge base, such as WordNet, brings with it certain difficulties arising from its structural and semantic idiosyncracies. Chapter 4, therefore, sets out an in-depth structural analysis of the topology of WordNet aimed at addressing these difficulties. Section 4.3 describes an experiment to correlate human judgments of concept specificity with topological features of the knowledge base which may then be exploited in constructing a representation of text. Chapter 5 discusses Storylines, a system designed to construct such representations from free text. The final representation described in section 5.8 is a rich graph structure whose latent information content includes text-based cohesive relations, WordNet-derived semantic relations and content and WordNet topological features. Chapter 6 describes how this structure has been exploited in the task of text differentiation which comprises the related tasks of similarity estimation and differential content detection. Chapter 7 describes a psycholinguistic experiment to derive a gold standard of human similarity judgments to be used in the evaluation of the Storylines similarity estimation component. This chapter provides an in-depth analysis of a range of measures of inter-rater agreement applied to the experiment data. Chapter 8 gives a comparative analysis of the results of the Storylines performance evaluation

on the gold standard data described in Chapter 7 and two other datasets, the Australian News Text corpus (see section 8.2.3) and the RTE Rel corpus (see section 8.2.2). This chapter also outlines a pilot evaluation of the Storylines differential content component.

In this final chapter, I draw together the threads of the thesis. Section 9.2 sets out new directions for future work suggested by this research. Sections 9.3 and 9.4 discuss the research contributions and conclusions of the thesis.

9.2 Future Work

The research reported in this thesis has opened up several areas which invite future work. Some of these are outlined below.

9.2.1 Lexical primes and seeding the representation

Section 2.6.2 outlined a challenging claim made by Hoey (2004) that the phenomenon that we call "lexical cohesion" is dictated by a process of "lexical priming", whereby particular lexical items prime particular other lexical items and that these connections are the governing principles of the mental lexicon and of text production and processing. This claim has some support in research on reading whereby it is reported that comprehension is facilitated if readers are primed on key content words. A very interesting direction for future development of the Storylines system involves incorporating this notion of priming into text representation construction. This would entail designing a new construction algorithm that begins with certain seed or primed concepts and builds the representation incrementally from these. This also provides a means of handling ambiguity, whereby the contribution of differ-

ent senses of ambiguous words may be evaluated prior to their addition to the representation. This approach is also appealing as it aims to model the *processes* of text comprehension in order to build a representation of text meaning.

9.2.2 Small World Graphs

As noted in section 4.2.9, the inclusion of polysemy relations between nodes in the WordNet hierarchy convert its low connectivity graph structure to a high connectivity "small world" network (Sigman and Cecchi 2002). The current implementation of Storylines encodes these relations between WordNet synsets but does not yet exploit its potential. "Small world" networks, described in (Watts and Strogatz 1999), are highly connected networks that exist throughout nature and society. In the context of a knowledge base, a small world network is organised around hubs of central concepts and the distance between nodes in the network is low with much traffic passing through these central hubs. This may indeed prove to be a better model of the mental lexicon than the usual hierarchical structure of ontologies. This structure presents great potential for the opportunist text representation algorithm. One interesting possibility is to address the question of word sense ambiguity through the inherent under–specification of this model. Ambiguities may, in fact, be collapsed in the central semantic hubs of the representation, obviating the need to select specific WordNet synsets for each word.[1]

[1]The problem of the uneven population of the WordNet database has been discussed at length in this thesis. The knowledge base also suffers from the related problem of over-fine granularity of concept definition within synsets. The central hubs of a small world network represent the possiblity of collapsing these over-fine distinctions for polysemous words.

9.2.3 Multiword Expressions and Proper Nouns

A more mundane extension of the Storylines system involves providing a more comprehensive analysis of multi-word expressions and proper nouns. In its current implementation, Storylines relies on the LT CHUNK tagger to identify any multiword expressions in the source text. Furthermore, there is no Proper Noun identification algorithm implemented in the system. These two areas constitute minor design changes for possibly major performance enhancement.[2]

9.2.4 Adding Logical and Rhetorical Relations

Chapter 2 discussed the multi-dimensional nature of text structure, from low-level lexical chains to high-level intentional structure. The Storylines implementation currently aims to model only the cohesive structure of text. Future development of the system should aim to address aspects of the coherent structure of text also. A first step in this direction should be inclusion of explicit logical or rhetorical relations. Negation, for example, would add a whole new semantic dimension to the text model which would be of particular importance in the differential information detection task. It should be noted that the research of Harabagiu (1999) would suggest that this is a far from trivial task.

[2]I fully acknowledge that the analysis of multiword expressions and proper nouns is very much an open topic in natural language processing and implementing such modules is a non-trivial task. They do, however, constitute minor changes in the sense that the inclusion of these modules does not entail a design overhaul of the Storylines system as it is currently implemented.

9.2.5 Skewed Similarity Experiment

Chapter 7 discussed the experiment I carried out to derive similarity judgments of text pairs. The texts for this experiment were selected at random from a corpus of RTÉ news texts. As a result, no identical pairs of texts were presented for comparison to the experiment subjects. It would be interesting to repeat this experiment with hand–selected texts in an attempt to deliberately skew the results towards similar judgments. The results could constitute a interesting point of comparison to the results of the experiment conducted and possibly throw further light on the notion of similarity as it is understood by humans.

9.2.6 Differential Content Output

As noted in section 8.6.3, the output of the differential information detection component of Storylines is very restrictive and should be extended. This issue could be addressed most profitably by empirical evaluation. I would propose a repeat of the psycholinguistic experiment described in section 8.6.1 on a larger scale, where subjects are requested to identify the most significant differences between text pairs. This corpus should be used, not as a gold standard for testing, but rather as a dataset for analysis of what kinds of differences in content people find important. This study could form the basis of a refined differential content component whose approach is empirically grounded and that exploits the latent information in the text representation graph more profitably.

9.2.7 Concept Specificity

A second area for further empirical investigation is the concept specificity metric. One approach for continued research in this area is to repeat the experiment, described in section 4.3, on a much larger scale and controlling for additional topological features to determine which of these features of WordNet are the best predictors of concept specificity in the knowledge base. Alternatively, or in parallel, an experiment could be conducted to validate the node specificity and normalised node specificity formulae, set out in section 6.6.1. The results of these formulae were evaluated informally for the research reported in this thesis. It would be interesting to validate this informal evaluation in a large-scale study.

9.2.8 Similarity *versus* Relatedness

Section 8.5.5 raised questions regarding the nature of text similarity and text relatedness. This section notes that similarity in text is a multi-dimensional concept that may or may not correlate with relatedness of text on one or many dimensions. A final direction for future work involves investigating what, in fact, are the dimensions of similarity as it is understood for texts and how or where does the notion of relatedness fit in to these dimensions. In essence, this is a theoretical issue that is closely related to the question of what is a text in general. In this context, an interesting approach could investigate whether people's notions of types of text similarity correlate with De Beaugrande and Dressler's (1981) standards of textuality.

9.3 Thesis Contributions

This section outlines the main contributions of this thesis. Some contributions fulfil the aims for this research set out in Chapter 1. Others emerged from an analysis of empirical findings in this research. The contributions are grouped under four headings: provision of computational tools, knowledge base structural analysis, validation and evaluation and theoretical findings.

The provision and development of computational tools for text representation and comparison

- Chapter 5 describes the design and implementation of a system, Storylines, to construct a meaningful representation of text within the framework of lexical cohesion that exploits text-based cohesive devices and the semantic content and structural relations of the WordNet knowledge base. This representation is constructed from free text without mark-up and without disambiguation, with the only preprocessing carried out by the LT CHUNK tagger and chunker (Mikheev 1996, Finch and Mikheev 1997). It uses all open class parts of speech, except adverbs as the inter-syntactic category connectivity for adverbs in WordNet is low.

- Chapter 6 describes the design and implementation of a similarity estimation algorithm which exploits all features of the afore-mentioned representation to determine the similarity of texts.

- Section 6.8 describes the design and implementation of a differential content detection algorithm which exploits all features of the aforementioned representation to identify the differential information content of texts.

An in-depth analysis of WordNet and an empirical investigation of concept specificity aimed at controlling exploitation of the knowledge base as a real world knowledge source.

- Chapter 4 elaborates a full structural analysis of WordNet. This analysis examines key features of the topology of the knowledge base structure and identifies the topological features that best characterise this structure.

- Section 4.3 outlines an empirical experiment to investigate the notion of concept specificity in the WordNet knowledge base and how this correlates with the afore-mentioned topological features.

- The results of this experiment and the structural analysis of WordNet guided the derivation of an empirically-grounded metric of concept specificity in the knowledge base described in section 6.6.1. This metric has been used to factor out the non-uniformity of the structure of WordNet and to evaluate the informational contribution of its individual concepts (see section 6.6.2).

- The structural WordNet analysis and relevant research in the literature inspired the elaboration of a weighting algorithm for the different types of WordNet relation described in section 6.6.1. This algorithm was validated on the basis of experiments run using the Storylines similarity detection component.

Validation and evaluation: provision of resources, encapsulation of evaluation methodologies and thorough evaluation of Storylines components.

- A key resource for empirical validation was provided by the compilation of a corpus of real news texts(see section 3.3.1). This corpus was used in the evaluation of the text representation and differentiation model. In fact, the full potential of the corpus was not exploited as it includes news sources, broadcast news scripts and internet news articles and constitutes an ideal dataset for a comparative analysis of genre and style within the news domain.

- Chapter 7 elaborates the derivation of two gold standard resources, one of human similarity judgments and the other of human relatedness judgments, both of which were used in the evaluation of the Storylines similarity estimation component.

- Section 7.5 provides a detailed encapsulation of the literature on experiment data reliability and inter-rater agreement with my gold standard experiment data as case study.

- Chapter 8 provides an analysis of the effects of concept specificity, relation type and text-based cohesive devices on the WordNet-based representation of text. These were evaluated singly and in combination as techniques in the performance of the Storylines system on the similarity estimation task.

- Thorough evaluation of the Storylines similarity estimation component using three gold standard datasets is set out in section 8.5.

- Section 8.6 outlines a pilot evaluation of the Storylines differential content component using a small corpus of Google News texts.

Theoretical import of empirical findings

- A final contribution is the formulation of a hypothesis of a complex relationship between text similarity, topic similarity and text relatedness which emerged from a comparative analysis of the performance of the Storylines similarity detection component on the three gold standard datasets used in the evaluation process.

9.4 Concluding Remarks

At the outset of this thesis, I stated that the central focus of the research has been *meaning* in text. The computational approach taken addressed both the representation of meaning and the differentiation of meaning in text, providing an implementation of possible factors that aid in tackling these core issues. In parallel, what has emerged, through the empirical studies conducted as part of this research, is a clear and constant reminder of the multidimensionality of text. This serves to emphasise that it is through the grounding of computational and theoretical models in such empirical human data that we, as researchers, are made aware of how far we have come and how far there is to go in our search for meaning.

Bibliography

Eneko Agirre and German Rigau, 1995. A proposal for word sense disambiguation using conceptual distance. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria.

Reem Al-Halimi and Rick Kazman, 1998. Temporal Indexing through Lexical Chaining. In *WordNet,an electronic lexical database*, pp. 333–351. The MIT Press.

Yaser Al-Onaizan, Ulrich Germann, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Daniel Marcu, and Kenji Yamada, 2000. Translating with Scarce Resources. In *The 17th National Conference of the American Association for Artificial Intelligence AAAI'2000*, Austin, Texas.

Laura Alemany and Maria Fuentes, 2003. Integrating cohesion and coherence for automatic summarization. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-03)*, Budapest, Hungary.

Nicholas Asher and Alex Lascarides, 1994. Intentions and information in discourse. In *Proceedings of the 32nd Annual Meeting of the Association*

for Compuational Linguistics (ACL-94), pp. 34–41, New Mexico State University, Las Cruces, New Mexico.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto, 1999. *Modern information retrieval.* Harlow: Addison-Wesley.

J. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, 1989. A Tree-Based Statistical Language Model for Natural Language Speech Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37-7 (1989), 1001–1008.

Regina Barzilay and Michael Elhadad, 1997. Using Lexical Chains for Text Summarization. In *Proceedings of ACL-97/EACL-97, Workshop on Intelligent Scalable Text Summarization*, pp. 10–17.

John A. Bateman and Christian M. I. M. Matthiessen, 1993. Uncovering the text base. In *Selected papers from the International Conference on Texts and Language Research*, Keqi Hao and Hermann Bluhme, (ed), Xi'an, china. Jiaotong University Press.

Robert de Beaugrande and Wolfgang Dressler, 1981. *Introduction to Text Linguistics.* London and New York, Longman.

Yvonne M. M. Bishop, Stephen E. Fienberg, and Paul W. Holland, 1975. *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Massachusetts and London, England, The MIT Press.

Nadjet Bouayad-Agha, Richard Power, and Donia Scott, 2000. Can text structure be incompatible with rhetorical structure? In *Proceedings of the International Conference in Natural Language Generation*, Mitzpe Ramon, Israel.

Thorsten Brants, 2000. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA.

Marten Brouwer, Cedric C. Clarke, George Gerbner, and Klaus Krippendorff, 1969. The television world of violence. In *Mass media and violence: A report to the National Commission on the causes and prevention of violence*, R. K. Baker and S. J. Balls, (ed), pp. 519–591. Government Printing Office, Washington, DC.

Gillian Brown and George Yule, 1983. *Discourse Analysis.* Cambridge, Cambridge University Press.

M. Brunn, Y. Chali, and C. J. Pinchak, 2001. Text Summarization Using Lexical Chains. In *Proceedings of the Document Understanding Conference (DUC-2001)*, pp. 135–140.

Alexander Budanitsky, 1999. Lexical Semantic Relatedness and its Application in Natural Language Processing. Technical Report CSRG-390, Department of Computer Science, University of Toronto.

Jill Burstein and Daniel Marcu, 2000. Towards Using Text Summarization for Essay-Based Feedback. In *Proceedings of Le 7e Conference Annuelle sur Le Traitement Automatique des Langues Naturelles TALN'2000*, Lausanne, Switzerland.

Jean Carletta, 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22-2 (1996), 249–254.

Joe Carthy, 2002. *Lexical Chains for Topic Tracking.* Phd. Thesis, Department of Computer Science, University College Dublin.

Joe Carthy and Alan F. Smeaton, 2000. The Design of a Topic Tracking System. In *Proceedings of the 22nd BCS-IRSG Colloquim on Information Retrieval Research.*

Y. Chali, M. Kolla, N. Singh, and Z. Zhang, 2003. The Univeristy of Lethbridge Text Summarizer at DUC-2003. In *Proceedings of the Document Understanding Conference (DUC-2003)*, pp. 148–152.

Freddy Y. Y. Choi, 2000. Advances in domain independent linear text segmentation. In *Proceedings of ANLP 2000*, pp. 26–33.

Freddy Y. Y. Choi, Peter Weimer-Hastings, and Johanna Moore, 2001. Latent Semantic Analysis for Text Segmentation. In *Proceedings of the 2001 Conference on Empirical Methods in NLP*, Lillian Lee and Donna Harman, (ed), pp. 109–117.

G. G. Chowdhury, 2004. *Introduction to Modern Information Retrieval.* London, Facet Publishing, tweede edition.

Kenneth Church and William Gale, 1995. Poisson mixtures. *Journal of Natural Language Engineering*, 1-2 (1995), 163–190.

William Gemmell Cochran, 1950. The comparison of percentages in matched samples. *Biometrika*, 37 (1950), 256–266.

Jacob Cohen, 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20 (1960), 37–46.

Wayne Cowart, 1997. *Experimental Syntax: Applying objective methods to sentence judgments.* Thousand Oaks, California, Sage Publications.

S Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41 (1990), 391–407.

Ann Devitt and Carl Vogel, 2003a. The Topology of WordNet: some metrics. In *Proceedings of the Second Global WordNet Conference 2004*, pp. 106–112, Brno, Czech Republic. Masaryk University, Brno.

Ann Devitt and Carl Vogel, 2003b. Using WordNet Hierarchies to Pinpoint Differences in Related Texts. In *Proceedings of EUROLAN 2003, Ontologies and Information Extraction International Workshop*, pp. 37–44, Bucharest, Romania.

Willian R. Dillon and Narendra Mulani, 1984. A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioural Research*, 19 (1984), 438–458.

J. Ellman, 2000. *Using Roget's Thesaurus to Determine the Similarity of Texts.* Phd. Thesis, Department of Computer Science, University of Sunderland.

N. E. Enkvist, 1978. Coherence, pseudo-coherence, and non-coherence. In *Cohesion and semantics*, J.-O. Östman, (ed). Abo Akademi Foundation, Abo, Finland.

Barbara di Eugenio and Michael Glass, 2004. The Kappa statistic: a second look. *Computational Linguistics*, 30-1 (2004), 95–101.

Christiane Fellbaum, 1998. *WordNet,an electronic lexical database.* The MIT Press.

David Fernández-Amorós, Julio Gonzalo, and Felisa Verdejo, 2001. The Role of Conceptual Relation in Word Sense Disambiguation. In *NLDB 2001*, pp. 87–98.

Steve Finch and Andrei Mikheev, 1997. A Workbench for Finding Structure in Texts. In *Proceedings of Applied Natural Language Processing*, Washington, D.C.

Peter W. Foltz, 1996. Latent Semantic Analysis for Text-Based Research. *Behavior Research Methods, Instruments and Computers*, 28-2 (1996), 197–202.

P. W. Foltz, W. Kintsch, and T. K. Landauer, 1998. The Measurement of Textual Coherence with Latent Semantic Analysis. *Discourse Processes*, 25 (1998), 285–307.

William Gale, Kenneth Ward Church, and David Yarowsky, 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of ACL92*, pp. 249–256, Newark, Delaware.

M. Galley and K. McKeown, 2003. Improving Word Sense Disambiguation in Lexical Chaining. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03).*

Michael R. Garey and David S. Johnson, 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* New York, W. H. Freeman and Company.

M. F. Garrett, 1982. Production of Speech: Observations from Normal and Pathological Language Use. In *Normality and Pathology in Cognitive Functions*, A. Ellis, (ed). Academic Press, London.

Emden R. Gasner, Eleftherios Koutsofios, Stephen C. North, and Kiem-Phong Vo, 1993. A technique for drawing directed graphs. *IEEE Transactions on Software Engineering*, 19-3 (1993), 124–230.

Dedre Gentner and Arthur B. Markman, 1994. Structural Alignment in Comparison: No difference without similarity. *Psychological Science*, 5-3 (1994), 152–158.

Stephen J. Green, 1999. Building Hypertext Links By Computing Semantic Similarity. *IEEE Transactions on Knowledge and Data Engineering*, 11-5 (1999), 713–730.

P. Grice, 1975. Logic and Conversation. In *Studies in Syntax*, P. Cole and J. L. Morgan, (ed), Vol. 3. Academic Press, New York.

Barbara J. Grosz and Candace L. Sidner, 1986. Attention, Intentions and the structure of discourse. *Computational Linguistics*, 12-3 (1986), 175–204.

William N. Grove, Nancy C. Andreasen, Patricia McDonald-Scott, Martin B. Keller, and Robert W. Shapiro, 1981. Reliability Studies of Psychiatric Diagnosis. Theory and Practice. *Archives of General Psychiatry*, 38 (1981), 408–413.

Michael A. K. Halliday and Ruqaiya Hasan, 1976. *Cohesion in English.* Longman.

Sanda M Harabagiu, 1999. From Lexical Cohesion to Textual Coherence: - A Data Driven Perspective. *Journal of Pattern Recognition and Artificial Intelligence*, 13-2 (1999), 247–265.

Sanda M Harabagiu and Steven J Maiorano, 1999. Knowledge-lean Coreference Resolution and its Relation to Textual Cohesion and Coherence. In

Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 29–38, University of Maryland. Workshop on the relation of discourse/dialogue structure and reference.

Sanda M Harabagiu, George A. Miller, and Dan Moldovan, 1999. WordNet 2: a morphologically and semantically enhanced resource. In *Proceedings of SIGLEX 99*, pp. 1–8, University of Maryland.

Sanda M. Harabagiu and Dan Moldovan, 1998. A parallel system for text inference using marker propogations. *IEEE Transactions in Parallel and Distributed Systems*, 9-8 (1998), 729–747.

Vasileios Hatzivassiloglou and J. M. Wiebe, 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th International Conference on Computational Linguistics*, New Brunswick, NJ.

Marti Hearst, 1997. Text-tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23-1 (1997), 33–64.

Claire Hewson, Peter Yule, Dianna Laurent, and Carl Vogel, 2003. *Internet Research Methods: A practical guide for the social and behavioural sciences.* New Technologies for social research. London, Sage Publications.

D. Hiemstra, 2001. *Using Language Models for Information Retrieval.* Phd. Thesis, Centre for Telmatics and Information Technology, University of Twente.

Graeme Hirst and David St-Onge, 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In *WordNet,an electronic lexical database*, pp. 305–332. The MIT Press.

Jerry R. Hobbs, 1990. *Literature and Cognition.* CSLI Lecture Notes 21. Stanford, CA., CSLI.

Jerry R. Hobbs, 1995. Why is Discourse Coherent. In *Coherence in spontaneous text*, M. A. Gernsbacher and T. Givón, (ed), *Typological Studies in Language*, Vol. 31, pp. 29–70. John Benjamins Publishing Company.

Michael Hoey, 1991. *Patterns of Lexis in text.* Oxford University Press.

Michael Hoey, 2004. Cohesion and Collocation: Two Aspects of the same Phenomenon. AESLA 2004, Universidad Politecnica de Valencia, Spain.

Eduard Hovy, 1991. Approaches to the planning of coherent text. In *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, Cécile Paris, William Swartout, and William Mann, (ed), pp. 83–102. Kluwer Academic Publishers, Boston.

Edouard H Hovy, 1993. Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63 (1993), 341–385.

Hongyan Jing, 2000. Sentence simplification in automatic text summarization. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP'00)*, Seattle, Washington.

Philip N. Johnson-Laird, 1983. *Mental Models.* Cambridge, Cambridge University Press.

Julia Jorgensen, 1990. The psychological reality of word senses. *Journal of Psycholinguistic Research*, 19 (1990), 167–190.

Hans Kamp and Uwe Reyle, 1993. *From Discourse to Logic: Introduction to Model-Theoretic Semantics of Natural Language, Formal Logic and*

Discourse Representation Theory, Studies in Linguistics and Philosophy, Vol. 42. London, Boston, Dordrecht, Kluwer Academic Publishers.

Jaap Kamps, Robert J. Mokken, Maarten Marx, and Maarten de Rijke, 2004. Using WordNet to measure semantic orientation of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Vol. IV, pp. 1115–1118, Paris. European Language Resources Association.

A. Kaplan and J. M. Goldsen, 1965. The reliability of content analysis categories. In *Language of Politics: Studies in quantitative semantics*, H. D. Laswell et al., (ed), pp. 83–112. MIT Press, Cambridge, MA.

Slava Katz, 1996. Distribution of content words and phrases in text and language modelling. *Journal of Natural Language Engineering*, 2-1 (1996), 15–60.

Stefan Kaufmann, 1999. Cohesion and collocation: Using context vectors in text segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 591–595, University of Maryland.

Frank Keller, Martin Corley, Steffan Corley, Lars Konieczny, and Amalia Todirascu, 1998. WebExp: A Java Toolbox for Web-Based Psychological Experiments. Technical Report HCRC/TR-99, Human Communication Research Centre, University of Edinburgh.

Sarah Kenny, 1998. A Generic Automatic Experiment Creation and Presentation Tool. CSLL Final Year Project, Trinity College Dublin.

Adam Kilgarriff, 2001. Comparing Corpora. *International Journal of Corpus Linguistics*, 6-1 (2001), 1–37.

Alistair Knott, 1991. *New strategies and constraints in RST-based text planning.* Masters Thesis, Department of Artificial Intelligence, University of Edinburgh.

Alistair Knott and Chris Mellish, 1996. A data-driven method for classifying connective phrases. *Journal of Language and Speech*, 39 (1996).

Hideki Kozima and A Ito, 1997. Context Sensitive Word Distance by Adaptive Scaling of a Semantic Space. In *Recent Advances in Natural Language Processing: Selected Papers from RANLP 1995*, R. Mitkov and N. Nicolov, (ed), *Amsterdam Studies in the Theory and History of Linguistic Science: Current Issues in Linguistic Theory*, Vol. 136, Chapter 2, pp. 111–124. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Klaus Krippendorff, 1980. *Content Analysis: an Introduction to its Methodology.* Beverly Hills, CA, Sage Publications.

Klaus Krippendorff, 2004. Computing Krippendorff's alpha-reliability. published on the author's web site.

R. Kuhn and R. De Mori, 1990. A Cache-Based Natural Language Model for Speech Reproduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12-6 (1990), 570–583.

J. Lafferty and C. Zhai, 2001. Document Language Models, Query Models and Risk Minimization for Information Retrieval. In *Proceedings of the 24st Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-01)*, pp. 111–119.

D. Laham, 1997. Latent Semantic Analysis approaches to categorization. In *Proceedings of the 19th annual meeting of the Cognitive Science Society*, M. G. Shafto and P. Langley, (ed), pp. 979. Erlbaum, Mawhwah, NJ.

Claudia Leacock and Martin Chodorow, 1998. Combining local context with WordNet similarity for word sense identification. In *WordNet,an electronic lexical database*, pp. 265–283. The MIT Press.

Michael D Lee, B M Pincombe, and M B Welsh, 2004. A comparison of machine measures of text document similarity with human judgments. Submitted.

Douglas B. Lenat and R. V. Guha, 1990. *Building large knowledge-based systems : representation and inference in the Cyc project.* Reading, Massachusetts: Addison Wesley.

John Locke, 1964. *An essay concerning human understanding.* Collins, 5 edition.

Inderjeet Mani and Eric Bloedorn, 1999. Summarizing Similarities and Differences Among Related Documents. *Information Retrieval*, 1 (1999), 35–67.

William C. Mann and Sandra A. Thompson, 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8-3 (1988), 243–281.

Daniel Marcu, 1997a. The rhetorical parsing of natural language texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain.

Daniel Marcu, 1997b. *The Rhetorical Parsing, Summarization and Generation of Natural Language Texts.* Phd. Thesis, Department of Computer Science, University of Toronto.

Daniel Marcu, 1999. Discourse trees are good indicators of importance in text. In *Advances in Automatic Text Summarization*, I. Mani and M. Maybury, (ed), pp. 123–136. The MIT Press.

Daniel Marcu, 2000. Extending a Formal and Computational Model of Rhetorical Structure Theory with Intentional Structures la Grosz and Sidner. In *Proceedings of the 18th International Conference on Computational Linguistics COLING'2000*, Saarbrucken, Germany.

M. Marcus, Beatrice Santorini, and M. A. Marcinkiewicz, 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19-2 (1993), 313–330.

Christian Matthiessen and Sandra A. Thompson, 1988. The structure of discourse and subordination. In *Clause Combining in Grammar and Discourse*, Sandra A. Thompson and Haiman, (ed). Benjamins.

Kathy R. McKeown, 1985. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text.* Cambridge University Press.

Julia Medori, 2004. *Experiments Testing the Reliability and Validity of Author Identification Methods.* Masters Thesis, Computer Science Department, Trinity College Dublin.

Rada Mihalcea and Dan Moldovan, 2000. An Iterative Approach to Word Sense Disambiguation. In *Proceedings of Florida Artificial Intelligence Research Society Conference (FLAIRS 2000)*, pp. 219–223, Orlando, FL.

Andrei Mikheev, 1996. Learning part-of-speech guessing rules from the lexicon. In *Proceedings of the 16th International Conference on Computational Linguistics (Coling 96)*, pp. 770–775, Copenhagen.

George A. Miller and Christiane Fellbaum, 1991. Semantic networks of English. *Cognition*, 41-1-3 (1991), 197–229.

E. Miltsakaki and K. Kukich, 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10-1 (2004), 25–55.

Johanna D. Moore and Cécile L. Paris, 1989. Planning text for advisory dialogues. In *Proceedings of the Twenty-Seventh Annual Meeting of the Association for Computational Linguistics (ACL-89)*, pp. 203–211, Vancouver, B.C.

Johanna D. Moore and Cécile L. Paris, 1992. Planning text for advisory dialogues: Capturing intentional, rhetorical and attentional information. Technical Report 92-22, University of Pittsburgh Computer Science Department, Pittsburgh, PA.

Johanna D. Moore and Martha E. Pollack, 1992. A Problem for RST: The Need for Multi-Level Discourse Analysis. *Computational Linguistics*, 18-4 (1992), 537–544.

Jane Morris and Graeme Hirst, 1991. Lexical cohesion computed by thesaural relations as an indication of the structure of text. *Computational Linguistics*, 17-1 (1991), 21–48.

Megan Moser and Johanna D. Moore, 1995. Using Discourse Analysis and Automatic Text Generation to Study Discourse Cue Usage. In *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Understanding and Generation*, pp. 92–98, Menlo Park, CA. AAAI Press.

M. Okumura and Y. Honda, 1994. Word Sense Disambiguation and Text Segmentation Based on Lexical Cohesion. In *Proceedings of the 15th In-*

ternational Conference on Computational Linguistics (COLING-94), pp. 775–761.

Barbara Partee, Alice ter Meulen, and Robert Wall, 1993. *Mathematical Methods in Linguistics.* Kluwer Academic Publishers.

Rebecca J. Passonneau and Diane J. Litman, 1997. Discourse Segmentation by Human and Automated Means. *Computational Linguistics*, 23-1 (1997), 103–140.

F. Peng and D. Schuurmans, 2003. Combining naive Bayes and n-gram language models for text classification. In *Proceedings of the Twenty-Fifth European Conference on Information Retrieval Research (ECIR-03).*

M. Philips, 1985. *Aspects of Text Structure: an investigation of the lexical organisation of text.* Amsterdam, North-Holland.

Livia Polanyi, 1988. A formal model of the structure of discourse. *Journal of Pragmatics*, 12 (1988), 601–638.

J. Ponte and W. B. Croft, 1998. A Language Modelling Approach to Information Retrieval. In *Proceedings of the 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98)*, pp. 275–281.

Richard Power, 2000. Mapping rhetorical structures to text structures by constraint satisfaction. Technical Report ITRI-00-01, ITRI, University of Brighton.

Uta E. Priss, 1998. The Formalization of WordNet by Methods of Relational Concept Analysis. In Fellbaum98 (Fellbaum 1998), pp. 179–196.

J. F. Quesada, W. Kintsch, and E. Gomez, 2002. A theory of Complex Problem Solving using Latent Semantic Analysis. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, W. D. Gray and C. D. Schunn, (ed), pp. 750–755, Fairfax, VA. Lawrence Erlbaum Associates , Mahwah, NJ.

Philip Resnik, 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI-95*, pp. 448–453, Montreal, Quebec. Morgan Kaufmann.

Philip Resnik, 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Applications to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11 (1999), 95–130.

T Rietveld and R van Hout, 1993. *Statistical Techniques for the study of language and language behaviour.* Mouton de Gruyter.

C. J. Van Rijsbergen, 1989. Towards an information logic. In *Proceedings of the 12th ACM SIGIR Conference*, pp. 77–86.

Lucia Helena Machado Rino and Donia R. Scott, 1996. A Discourse Model for Gist Preservation. Technical Report ITRI-96-12, Information Technology Research Institute, University of Brighton.

P. Roget, 1977. *Roget's International Thesaurus*, Vol. 4. Harper and Row Publishers Inc.

Mark Rorvig, 1999. Images of similarity: a visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. *Journal of the American Society for Information Science*, 50-8 (1999), 639–651.

Dietmar Rosner and Manfred Stede, 1992. Customizing RST for the automatic production of technical manuals. In *Proceedings of the Sixth International Workshop on Natural Language Generation*, pp. 199–215, Berlin Heidelberg, Germany. Springer-Verlag.

Sheldon Ross, 1976. *A First Course in Probability.* Macmillan.

Paolo Rosso, Francesco Masulli, and Davide Buscaldi, 2003. Word Sense Disambiguation combining conceptual distance, frequency and gloss. In *Proceedings of the 2003 International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, China. IEEE.

Gerard Salton and Michael J. McGill, 1983. *Introduction to modern information retrieval.* New York, London, McGraw-Hill.

Howard Schumann and Stanley Presser, 1981. *Questions and answers in attitude surveys: Experiments on question form, wording and context.* New York, NY, New York Academic Press.

E Shipstone, 1960. Some variables affecting pattern conception. *Psychological monographs, general and applied*, 74 (1960), 1–41.

Sidney Siegel and N John Castellan, 1988. *Nonparametric statistics for the behavioral sciences.* Boston, MA, McGraw Hill.

Mariano Sigman and Guillermo A. Cecchi, 2002. The Global Organization of the Wordnet Lexicon. *Proceedings of the National Academy of Sciences of the USA*, 99-3 (2002), 1742–1747.

H. Gregory Silber and Kathleen F. McCoy, 2002. Efficiently Computed Lexical Chains As an Intermediate Representation for Automatic Text Summarization. *Computational Linguistics*, 28-4 (2002), 487–496.

David Singleton, 2000. *Language and the Lexicon: an Introduction.* Arnold, London.

Steven S. Skiena, 1997. *The Algorithm Design Manual.* New York, Springer-Verlag.

Andrew E. Smith, 2000. Machine Mapping of Document Collections: the Leximancer System. In *Proceedings of the Fifth Australian Document Computing Symposium*, Sunshine Coast, Australia. DSTC.

Mark A. Stairmand, 1996. *A Computational Analysis of Lexical Cohesion wtih Applications in Information Retrieval.* Phd. Thesis, Department of Language Engineering, University of Manchester Institute of Science and Technology.

Mark A. Stairmand, 1997. Textual context analysis for information retrieval. In *Proceedings of the 20th Annual ACM SIGIR Conference on Research and Development in IR*, pp. 140–147.

Nicola Stokes, 2004. *Applications of Lexical Cohesion Analysis in the Topic Detection and Tracking Domain.* Phd. Thesis, Department of Computer Science, University College Dublin.

Nicola Stokes and Joe Carthy, 2001. Using Data Fusion to Improve First Story Detection. In *Proceedings of the 23rd BCS IRSG European Conference on Information Retrieval Research*, pp. 78–90.

Nicola Stokes, Joe Carthy, and Alan F. Smeaton, 2002. Segmenting Broadcast News Streams using Lexical Chaining. In *Proceedings of STAIRS 2002*, pp. 145–154, Lyons, France.

Michael Sussna, 1993. Word sense disambiguation. In *Proceedings of the second international conference on information and knowledge management (CIKM-93)*, pp. 67–74, Arlington, Virginia.

David Touretsky, 1986. *The Mathematics of Inheritance Systems.* Los Altos CA, Morgan Kaufman.

Allen B. Tucker, Sergei Nirenburg, and Victor Raskin, 1986. Discourse and cohesion in expository texts. In *Proceedings of the 11th International Conference on Computational Linguistics*, pp. 181–183, Bonn, Germany.

Peter D. Turney, 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pp. 417–424, Philadelphia, Pennsylvania.

P. D. Turney and M. L. Littman, 2003. Learning Analogies and Semantic Relations. Technical Report ERB-1103, National Research Council, Institute for Information Technology.

Amos Tversky, 1977. Features of Similarity. *Psychological Review*, 84-4 (1977), 327–351.

John S. Uebersax, 1999. Probit latent class analysis: conditional independence and conditional dependence models. *Applied Psychological Measurement*, 23-4 (1999), 283–297.

John S. Uebersax, 2004. Statistical Methods for Rater Agreement. published on the author's website, downloaded June 2004.

D.J. Watts and S.H. Strogatz, 1999. Collective dynamics of small world networks. *Nature*, 401-130 (1999).

Peter Weimar-Hastings, Karen Wiemar-Hastings, and A. Graesser, 1999. Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In *Artificial Intelligence in Education*, S. Lajoie and M. Vivet, (ed), pp. 535–542. IOS Press, Amsterdam.

Janyce Wiebe, Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane Litman, David Pierce, Ellen Riloff, Theresa Wilson, David Day, and Mark Maybury, 2003. Recognizing and Organizing Opinions Expressed in the World Press. In *AAAI Spring Symposium on New Directions in Question Answering.*

Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara, 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association forComputational Linguistics (ACL-99)*, pp. 246–253.

Oscar Wilde, 1908. *The Picture of Dorian Gray.* Liepzig, Bernhard Tauchnitz.

M. E. Winston, R. Chaffin, and D. J. Herrmann, 1987. A Taxonomy of Part-Whole Relations. *Cognitive Science*, 11 (1987), 417–444.

M. B. Wolfe, M. E. Schreiner, B. Rehder, D. Laham, P. W. Foltz, W. Kintsch, and T. K. Landauer, 1998. Learning from text: Matching readers and text by Latent Semantic Analysis. *Discourse Processes*, 25 (1998), 309–336.

D. Yarowsky, 1995. Unsupervised word-sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association of Computational Linguistics (ACL-95)*, pp. 189–196, Cambridge, MA.

George Kingsley Zipf, 1935. *Psycho-Biology of Languages.* Houghton-Mifflin.

Appendix A

Corpora

A.1 RTE Story topics

1. Arson in Co. Tyrone
2. Teachers Unions dispute
3. British-Irish council
4. Bali bombing investigation
5. Cork drugs bust
6. Firefighters strike in NI
7. Air France hijack (unfounded story)
8. Flood tribunal - Dublin Cllrs implicated
9. Donegal Garda suspension appeal (Frank Shortt case)
10. Israeli incursions into Gaza
11. Kuwaiti soldier arrest for US soldier shooting
12. Morris tribunal terms of reference
13. Partnership for Peace and NATO meeting
14. Miss World in Nigeria riots
15. Post British-Irish council reports
16. Report on the sinking of the Pisces
17. Child pornography arrest
18. IMO recommentations on ports of refuge
19. Prestige tanker oil slick
20. Human cloning claims
21. Rape in Cork city
22. Stabilising the Euro
23. Fire in Bangor - woman died
24. Arrest for firebomb attack
25. Fishing industry and the Irish Box
26. Flooding in Cork and Kerry
27. Re-opening of the Forum for Peace and REconciliation

28. Weapons inspection in Iraq
29. 2 Palestinians killed in West Bank
30. Amnesty refused access to Irish prsinos for review
31. France and Spain oil tanker restrictions
32. Garda car rammed in robbery
33. Baby born in Garda station
34. The Budget - expectations, reports, responses
35. Concorde investigation after flight glitch
36. Violence in East Timor
37. Arson in Derry
38. Bombs found at Ikea Amsterdam
39. Iraq weapons inspections - US response
40. 2 Palestinians shot in Hebron
41. Blairs comments on the North
42. Pub robbery in Dublin
43. French exercise oil tanker restriction
44. William Trevor gets knighthood
45. Turkey and USA do deal on airbases
46. Bushfires in Australia - break out,
47. Solar eclipse in Southern hemisphere
48. Flood tribunal: appeal of late Cllr Hand's family rejected
49. Funeral of Jim Mitchell
50. Moriary tribunal: Lowry and O'Reilly met at the Curragh
51. Failed case for failure of pyramid scheme
52. New Irish tourism authority chief
53. Drugs find and arrest in Waterford
54. Two South African's charged for drugs crime
55. Survey says Anti-US feeling high

56. Cork gas explosion
57. Stowaways on Ghana flight died
58. Flood tribunal: Enda Conway's submission
59. Cork GAA footballers strike
60. Greenpeace protest against single-hull oil tanker
61. Speculation over European Interest rates
62. Iraq weapons inspectors accused of spying
63. Pakistani consulate in Macedonia bombed
64. Israel: Sharon plans for Palestinian state
65. Moriarty tribunal: Denis O'Brien gave misleading financial reports
66. European constitution says Prodi
67. Arrest in Cork Rape case
68. Shooting in North Belfast
69. Turkey's progress for EU accession
70. Car Tax to be raised
71. UK Catholic Church admit mistakes in handling allegations
72. Jeanie Johnston's new captain
73. MacDonalds bomb in Jakarta
74. 9 Palestinians killed in Gaza Strip
75. Murder investigation in Kilkenny
76. Plane crashes into Federal Reserve Bank in Miami
77. Release of man in Cork rape case
78. Robbery in Offaly

A.2 RTE Gold Standard Experiment Texts

Story 1: 271102_AMDIraq8

1. UN weapons inspectors return to Iraq
2. UN arms inspectors return to Iraq

United Nations weapons inspectors have resumed inspections in Iraq under the terms of a strongly worded Security Council resolution. A team of seventeen experts have carried out the first checks on suspected arms sites in nearly four years. They will be joined by a further 300 inspectors over the coming months. The UN Secretary General, Kofi Annan, has underlined that military action against Iraq is avoidable if President Saddam Hussein honours his commitments and co-operates fully with the inspectors. Mr Annan said that if Iraq acted as it should, the arguments in favour of war would be considerably weakened.

Story 2: 041202TM1745_AMDBudgetSources25

BUDGET - conclusion

Finance Minister Charlie McCreevy said today's budget is a prudent and planned response to the impact of the continued international downturn on our public finances. 'This is a Budget rooted in a vision of an Ireland which values stability as the necessary base for advancing the welfare of all its people', he said. 'It is a Budget which avoids the mistakes of the past and provides the foundation for renewed growth and development. This is a balanced and fair Budget, which responds properly to the international downturn by preventing the huge deficits of the past, but also by making sure that important areas like social welfare, pensions and health are prioritised in difficult times,' he said. 'This is Budget which will protect our long-term future and secure the welfare of all sections of our society,' McCreevy concluded.

Story 3: 041202TM1745_AMDBudgetSources14

budget - social welfare

Finance Minister Charlie McCreevy told the Dail that the total cost of Social Welfare improvements in today's Budget is E501 million in 2003 and E530 million in a full year. He said the maximum weekly personal rate for all old age pensions will be increased by E10 from the first week of January 2003. Child Benefit will be increased by E8 a month for each of the first and second children to E125.60 a month and by E10 for each of the third and subsequent children to E157.30 a month. The increases will be effective from next April and will cost E78.62 million in 2003 and E104.83 in a full year. From June 2003, the Respite Care Grant will be increased by E100 to E735.

Story 4: 061202_AMDMurder4

1. Gardai begin Kilkenny murder hunt
2. Gardai begin Kilkenny murder hunt

Gardai in Kilkenny have begun a murder investigation following the discovery of a woman's body in a house in the city. The 36-year-old mother of two was found by emergency services responding to reports of a fire at the house. At around 5.00pm yesterday evening

the emergency services were called to a reported fire at a semi-detached house in Greenfield Road, Kilkenny City. The body of a woman was taken from the house and brought to St Luke's Hospital. It is understood the woman had extensive injuries to her head and torso. Gardai have now launched a full-scale murder investigation, and a post mortem examination is being carried out by the State Pathologist, Dr John Harbison. Members of the Garda Technical Bureau have begun examining the scene, which remains cordoned off.

Story 5: 041202TM1513_AMDPyramid2

1. Pyramid scheme charges dismissed
2. Pyramid scheme: charges dismissed

The Dublin District Court has dismissed the case against a woman being sued over a Pyramid scheme. Lynda Laird was being sued by two investors, who claimed she had frustrated their efforts to receive a promised E45,000 in two months. Siblings Rose and Dermot Smedker said they invested E5,713 in the scheme. Judge O'Leary said they had accused Ms Laird of operating a scam but said, if that was the case, the plantiffs had taken part in the same scam. He said the scheme appeared to be a form of Pyramid investment but it fell outside the scope of the Pyramid Act. He said it was clear that the scheme worked on the basis that an early minority investing would win and a late majority would lose. He ruled against awarding any costs.

Story 6: 271102_AMDFloods6

1. Roads in south west hit by floods

Heavy overnight rain has made driving conditions hazardous in parts of the south west. The Ballyhooley Road along with the Ballyvolanea and Togher areas of Cork city are flooded. There are also floods on the M20 immediately outside Mallow, on the Limerick side at Kilknockin as well as on the back road between the Glanmire Road and Gleshaboy Waterworks. Many houses in these areas are also flooded, as are some in Blackpool in Cork. A high tide at 10.35am this morning may affect the city centre. The main Limerick to Cork road is partly flooded as are back roads in the Parteen area of Limerick city. Driving conditions on some roads in Kerry have been made dangerous by surface water.

Story 7: 041202TM1745_AMDBudgetReaction5

1. FG critical of new budget
2. 2003 budget is met with criticism

Fine Gael's Finance spokesman, Richard Bruton, has described the budget as 'genetically flawed'. He criticised the Minister for Finance for spending ten times the amount of revenue available. He said that the Minister had failed to face down those in his party who wanted to put electoral gain before the national interest. Fine Gael's Denis Naughten said people on social welfare were going to be less well off after this budget. This view was echoed by Audrey Deane of the Society of St Vincent de Paul, who said this was not a fair budget. Ms Deane said the E6 a week increase in basic Social Welfare payments was not enough. She added "it is quite shaming."

Story 8: 221102_AMDflood9

FLOOD

The Flood Tribunal has heard that an Independent Dublin councillor, John O'Halloran, has admitted receiving IEP 2,500 from the former lobbyist, Frank Dunlop, during the 1996 Dublin West By-election. In a statement being read at the enquiry, Mr O'Halloran said he personally solicited this money as a political donation and that it was not connected with any rezoning motion or planning matter. The Tribunal heard Brian O'Halloran - one of the former owners of the Carrickmines site - say he was never approached for money by any politician for money in relation to having the land rezoned, nor did he offer any inducement. Two other co-owners, Gerard Kilcoyne and Professor Austin Darragh, also said that they said they were never approached for money by any politician in relation to having the land rezoned nor did they offer any inducement.

Story 9: 041202TM1513_AMDEclipse3

1. Rare total solar eclipse in southern hemisphere
2. Total solar eclipse S. hemisphere

Astronomers and onlookers in southern Africa and Australia were treated to a rare total solar eclipse today. The moon passed between the sun and earth at 0515 GMT in Angola, spreading to parts of Namibia, Botswana, Zimbabwe and South Africa. Day turned into night in southern Australia at 0910 GMT, as the shadow spread at a rate greater than 5,000 kilometres per hour. Thousands of spectators flocked to the seaside town of Ceduna, which became eclipse central and mounted a 'solarbration' for the event. Amateur and professional observers labelled the spectacle of totality 'amazing'. The eclipse lasted 32 seconds in southern Australia.

Story 10: 041202TM1745_AMDMoriarty3

MORIARTY

The Tribunal has also heard of a meeting that took place in August 1995 between the auctioneer Mark Fitzgerald and Denis O'Brien at the Shelbourne hotel. At the time Mr Fitzgerald was organising a golf classic fundraiser for Fine Gael. Denis O'Brien told Mr Fitzgerald that he was facing an uphill struggle against Motorola in relation to the licence and he said there were rumours that Albert Reynolds was in line for a payoff if Motorola won the competition. The Tribunal has stressed that it has no basis for thinking that Albert Reynolds had acted in any way inappropriately. Denis O'Brien agreed to sponsor the fundraiser to the tune of IEP 4000. Mark Fitzgerald says he was unhappy with the fact that Mr O'Brien was a sponsor.

Story 11: 051202_AMDMoriarty2

1. More O'Brien evidence at Moriarty
2. More O'Brien evidence at Moriarty

The Moriarty Tribunal has heard that Denis O'Brien appears to have given misleading information regarding his consortium's finances to the team assessing bids for the second

mobile phone licence. At a formal presentation of their bid in September of 1995 Denis O'Brien told assessors that his consortuim had been guaranteed a loan worth 30m from the financial group Advent. He said this was an 'irrevocable commitment'. However, the Tribunal has presented documents which show that Advent had made no such commitment and that Denis O'Brien was aware of that. The tribunal has said it is 'unclear' how Denis O'Brien and other members of his presentation team came to make such statements.

Story 12: 041202TM1537_AMDTourism2

1. New tourism body established
2. New tourism authority established

The first head of a new tourism body established under the terms of the Good Friday agreement has been named. Shaun Quinn will be chief executive of the National Tourism Development Authority that will form part of the Tourism Ireland agency. He will take up the position in time for the start of the 2003 tourism season. The authority is to take over responsibility for developing the industry from Bord Failte, which will now concentrate on specialist functions, including sporting events such as the 2006 Ryder Cup golf. The NTDA will operate under the name Failte Ireland. Legislation to set it up was introduced last month in the Dail. Mr Quinn, who is currently in charge of the CERT tourism trade training body, was appointed by Tourism Minister John O'Donoghue.

Story 13: 041202TM1104AMDIraq5

1. US asks UN to step up arms inspections
2. US wants more UN pressure on Iraq

The US Defence Secretary, Donald Rumsfeld, has said any country with an intelligence service knows that Iraq possesses weapons of mass destruction. This follows Baghdad's repeated claim that it does not have any such weapons. It is understood the Bush administration has asked the United Nations to step up the pace of arms inspections to increase the pressure on Iraq. The UN Secretary General, Kofi Annan, has said he is pleased with the Iraqi government's co-operation with arms inspectors so far. He was speaking after Iraq announced that it would be sending a declaration of its weapons programme to the UN on Saturday, one day ahead of schedule.

Story 14: 051202_AMDAntiUS3

1. Global poll shows increased anti-Americanism
2. Global survey on anti-Americanism

A large survey of global public opinion has found evidence of increased anti-Americanism. However, it also revealed that most people thought the spread of science and technology from the United States was beneficial. The study by a Washington-based research committee, the Pew Global Attitudes Survey, involved 38,000 interviews in 44 countries and was headed by the former US Secretary of State, Madeleine Albright. The anti-American trend was most dramatic in muslim societies, such as Egypt and Pakistan. One of the research team, Bruce Stokes, said that although American culture appeared to be popular, US government policy was attracting growing criticism.

Story 15: 041202TM1745_AMDBudgetSources49

The Minister for Finance, Charlie McCreevy, has announced public spending increases in this afternoon's budget, but he firmly warned at the outset of his speech that the economic situation had worsened because of the weakness of the global economy. There will be public spending increases, but there are also increases in indirect taxation. While income tax remains unaffected, stamp duties are to go up significantly from midnight, with a large increase on the rate charged on credit cards. Duties on ATM cards go up to ten euro a year, and the duty on combined ATM/laser cards will be 20 euro a year. Petrol and beer have been spared, but the price of cigarettes, diesel, spirits and alcopops goes up. Fine Gael's Finance spokesman, Richard Bruton, described the budget as ğenetically flawed:

Story 16: 271102_AMDIsrael8

1. 2 Palestinian commanders in Jenin

A suspected Palestinian militant was killed in the Gaza Strip when the car he was driving, which was packed with explosives, blew up as it headed towards an Israeli military post. Elsewhere two Palestinian militant commanders were killed after an Israeli air to ground missile was fired at a house in the Jenin refugee camp in the West Bank. According to the reports, one of the men was a regional commander of the al-Aqsa Martyrs Brigades while the other was a leading member of Hamas. It has also been reported that Israeli troops shot dead a Palestinian near the West Bank city of Nablus overnight. In Gaza, two Israeli helicopter gunships are reported to have carried out a raid on western Khan Yunis in the southern Gaza Strip, badly damaging a local school.

Story 17: 051202_AMDProdi7

1. Prodi wants new Commission powers

EU Commission President Romano Prodi has proposed a federalist constitution, seeking wide new powers for the Commission and rejecting an appointed 'president of Europe'. In proposals to the Convention drafting a constitution for an enlarged EU, Prodi called for future Commission presidents to be elected by the European Parliament. 'We must build the first true supranational democracy in the world,' he told the European Parliament. Prodi proposed a single EU foreign minister who would sit in the Commission but be answerable both to member states and the Commission president. The Commission called for an extension of majority voting to almost all areas of EU policy except defence. Prodi also rubbished the idea of appointing a president of the European Council.

Story 18: 061202_AMDMacDoBomb2

1. Indonesian MacDonalds bomb kills three
2. Bomb at a MacDonald's kills three

A bomb at a McDonald's restaurant in Indonesia has killed three people and injured 11. Police are still trying to determine whether the Makassar blast yesterday evening is related to the October 12 Bali bombing and other incidents. The bomb went off in the kitchen at

the McDonald's in the city's Ratu Indah shopping mall. A second bomb exploded shortly afterwards at a Toyota showroom in the city in South Sulawesi province and damaged at least four cars. No one was hurt. A police source said both the showroom and the shopping mall were owned by top welfare minister Yusuf Kalla, a native of Makassar. Kalla negotiated an agreement last December to end Muslim-Christian battles in Central Sulawesi, and also a deal in February to end sectarian fighting in the Maluku islands.

Story 19: 221102_AMDpisces5

1. Pisces report is sent to families
2. Pisces report is sent to families

Copies of the report of the investigation into the sinking of the fishing vessel, the Pisces, have been sent to the families involved. Five people, including a 14-year-old boy, drowned in July of this year when the Pisces sank off the coast of Wexford. It later emerged that no one on board was wearing a lifejacket. The families will now review it with their legal advisors before making any submissions to the Marine Casualties Investigations Board. A final report will then be published. In August of this year a consultation process on the wearing of lifejackets was initiated by the Minister of the Marine Dermot Ahern.

Story 20: 271102_AMDFloods4

FLOODING

Cork is one of the worst areas affected by the overnight rain. There is flooding on the Ballyhooley Road and in the Ballyvolane and Togher areas of the city. There is also flooding on the N20 immediately outside Mallow, on the Limerick side at Kilknockin as well as on the back road between the Glanmire Road and Glashaboy Waterworks. Many houses in these areas are also flooded, as are some in Blackpool in Cork City. Gardai in the city have appealed to motorists to slow down. They also say the City Centre is still clear of flooding but high tide is at 10.35 and conditions could worsen. The main Limerick to Cork road also has flooding. And in the Parteen area of Limerick City, back roads have been flooded. In Kerry, no roads are impassible but water on roads is making driving conditions dangerous in places.

Story 21: 041202TM1745_AMDBudgetSources17

BUDGET-CARDS/BANKING

Today's budget has dramatically increased the stamp duty charged on cheques and credit cards. The stamp duty on each cheque will rise from 8 cent to 15 cent. The stamp duty on a credit card will rise from 19 euro to 40 euro per annum. The stamp duty on ATM cards will rise from 6.25 to 10 euro, whilst ATM cards with a Laser function will rise from 6.25 to 20 euro per card per annum. It's estimated these changes will yield over 50m euro to the exchequer next year. The minister for finance is levying a new 100m euro per annum tax on the banking and financial sector for the next three years. This tax will take the form of a special stamp duty on every financial company. It will be calculated based on the level of DIRT tax the bank paid in 2001, and the minister estimates it will be equal to 50% of this payment.

Story 22: 221102_AMDflood8

1. Flood hears councillors' denials
2. Cash for votes denied at Tribunal

The Flood Tribunal has heard formal written denials from two Dublin County Councillors that they accepted money in return for voting in favour of rezoning motions. Councillor Tony Fox denied ever receiving any money from Frank Dunlop. Councillor Colm McGrath said he was given a number of unconditional political donations from the lobbyist, but these were not connected with the planning process. Both men were named yesterday in an opening statement to the Tribunal as accepting bribes in return for their votes. It emerged at today's session that at one stage, the Tribunal chairman, Mr Justice Fergus Flood, wrote to the former Fine Gael Senator Liam Cosgrave, accusing him of failing to adequately address matters of interest to the Tribunal.

Story 23: 221102_AMDfirefighters8

1. Strike by NI firefighters under way
2. Firefighters' strike is under way

The British Army has taken over firefighting duties in the North at the start of an eight-day strike by fire crews. Almost 2,000 full-time firefighters have left their posts and pickets are being put on fire stations. The pay strike by firefighters throughout Britain and Northern Ireland began at 9am this morning after overnight talks failed to avert it. The Fire Brigades' Union has accused the British government of vetoing an agreement that could have prevented the action. The union said it had worked out a deal with employers that would have given firefighters a 16% pay increase, but the government had refused to help fund it.

Story 24: 051202_AMDFlood2

1. Frenxy of pressure on Councillors: witness
2. Flood: 'frenzy' pressure on Cllrs

A former senior planner with Dublin County Council has described how councillors were subjected to a frenzy of external pressures when making decisions on development plans. Enda Conway said on many occasions throughout the 1990s councillors went against all professional advice and made many 'illogical and unjustifiable' decisions, rezoning land to the benefit of single land owners. Mr Conway said it often resulted in bad planning decisions which incurred a cost to the community. He said planners and council officials found it an unpleasant and disappointing experience to have their advice and work set aside by councillors.

Story 25: 271102AMDFirebomb4

1. Man due in court over firebomb attempt in North
2. Man due in court for bomb attempt

A 34-year-old man is due in court today in the North charged in connection with an attempted firebomb attack in Belfast city centre at the weekend. The man has been

charged with possession of explosives with intent to endanger life and conspiracy to cause an explosion. He was arrested in a major security operation in the city centre last Sunday night, during which a second man was shot twice by police. Police discovered a device consisting of two gas cylinders and two pipe bombs linked to a number of containers of flammable liquid in a car abandoned outside the motor tax office at Upper Queen Street at the weekend. A timer device had been attached and activated. The shooting of the second man by police is being investigated by the North's Police Ombudsman, Nuala O'Loan.

A.3 Australian Gold Standard Experiment Texts

Story 1: OzText1

The national executive of the strife-torn Democrats last night appointed little-known West Australian senator Brian Greig as interim leader - a shock move likely to provoke further conflict between the party's senators and its organisation. In a move to reassert control over the party's seven senators, the national executive last night rejected Aden Ridgeway's bid to become interim leader, in favour of Senator Greig, a supporter of deposed leader Natasha Stott Despoja and an outspoken gay rights activist.

Story 2: OzText2

Cash-strapped financial services group AMP has shelved a $400 million plan to buy shares back from investors and will raise $750 million in fresh capital after profits crashed in the six months to June 30. Chief executive Paul Batchelor said the result was "solid" in what he described as the worst conditions for stock markets in 20 years. AMP's half-year profit sank 25 per cent to $303 million, or 27c a share, as Australia's largest investor and fund manager failed to hit projected 5 per cent earnings growth targets and was battered by falling returns on share markets.

Story 3: OzText3

The United States government has said it wants to see President Robert Mugabe removed from power and that it is working with the Zimbabwean opposition to bring about a change of administration. As scores of white farmers went into hiding to escape a round-up by Zimbabwean police, a senior Bush administration official called Mr Mugabe's rule "illegitimate and irrational" and said that his re-election as president in March was won through fraud. Walter Kansteiner, the assistant secretary of state for African affairs, went on to blame Mr Mugabe's policies for contributing to the threat of famine in Zimbabwe.

Story 4: OzText4

A radical armed Islamist group with ties to Tehran and Baghdad has helped al-Qaida establish an international terrorist training camp in northern Iraq, Kurdish officials say. Intelligence officers in the autonomous Kurdish region of Iraq told the Guardian that the Ansar al-Islam (supporters of Islam) group is harbouring up to 150 al-Qaida members in a string of villages it controls along the Iraq-Iran border. Most of them fled Afghanistan after the US-led offensive, but officials from the Patriotic Union of Kurdistan (PUK), which controls part of north-east Iraq, claim an "abnormal" number of recruits are making their way to the area from Jordan, Syria and Egypt.

Story 5: OzText5

Washington has sharply rebuked Russia over bombings of Georgian villages, warning the raids violated Georgian sovereignty and could worsen tensions between Moscow and Tbilisi. "The United States regrets the loss of life and deplores the violation of Georgia's sovereignty," White House spokesman Ari Fleischer said. Mr Fleischer said US Secretary

of State Colin Powell had delivered the same message to his Russian counterpart but that the stern language did not reflect a sign of souring relations between Moscow and Washington.

Story 6: OzText6

A gay former student of a Melbourne Christian school is taking legal action under equal opportunity legislation, claiming the school discriminated against him because of his sexuality. Tim, 16, alleged a staff member at Hillcrest Christian College in Berwick told him he "had the devil in him", and constant bullying by students prompted the principal to tell him to hide his sexuality. He left the school several weeks ago and is continuing Year 10 by distance education after he said homophobic bullies threw rocks at his head, spat on him, called him names and slashed his belongings.

Story 7: OzText7

Senior members of the Saudi royal family paid at least $560 million to Osama bin Laden's terror group and the Taliban for an agreement his forces would not attack targets in Saudi Arabia, according to court documents. The papers, filed in a $US3000 billion ($5500 billion) lawsuit in the US, allege the deal was made after two secret meetings between Saudi royals and leaders of al-Qa'ida, including bin Laden. The money enabled al-Qa'ida to fund training camps in Afghanistan later attended by the September 11 hijackers. The disclosures will increase tensions between the US and Saudi Arabia.

Story 8: OzText8

Palestinian hired gun Abu Nidal, whose violent death was reported last week from Baghdad, was murdered on the orders of Iraqi President Saddam Hussein after refusing to train al-Qa'ida fighters based in Iraq, reports said yesterday. Iraqi intelligence chief Taher Jalil Habbush said last Wednesday Abu Nidal had shot and killed himself after being discovered living illegally in Baghdad and facing interrogation for anti-Iraqi activities. But Western diplomats believe the radical militant was killed for refusing to reactivate his international terrorist network.

Story 9: OzText9

Hunan province remained on high alert last night as thunderstorms threatened to exacerbate the flood crisis, now entering its fifth day and with 108 already dead and hundreds of thousands evacuated. On the flood frontline at Dongting Lake, the water level peaked at just under 35m on Saturday night, then eased about 3cm during the day under a hot sun, with temperatures reaching 35C. But with the lake still brimming at dangerously high levels, and spilling over the top of its banks in some places, locals were fearful that a thunderstorm and high winds forecast to hit the region last night would damage the dikes. About 1800km of dikes around the lake are all that stand between 10 million people in the surrounding farmland and disaster.

Story 10: OzText10

A U.S.-British air raid in southern Iraq left eight civilians dead and nine wounded, the Iraqi military said Sunday. The military told the official Iraqi News Agency that the warplanes bombed areas in Basra province, 330 miles south of Baghdad. The U.S. Central Command in Florida said coalition aircraft used precision-guided weapons to strike two air defense radar systems near Basra "in response to recent Iraqi hostile acts against coalition aircraft monitoring the Southern No-Fly Zone."

Story 11: OzText11

Iraq and Russia are close to signing a $40 billion economic cooperation plan, Iraq's ambassador said Saturday, a deal that could put Moscow at odds with the United States as it considers a military attack against Baghdad. The statement by Ambassador Abbas Khalaf came amid indications that Russia, despite its strong support for the post-Sept. 11 antiterrorism coalition, is maintaining or improving ties with Iran and North Korea, which together with Iraq are the countries President Bush has labeled the "axis of evil."

Story 12: OzText12

U.S. intelligence cannot say conclusively that Saddam Hussein has weapons of mass destruction, an information gap that is complicating White House efforts to build support for an attack on Saddam's Iraqi regime. The CIA has advised top administration officials to assume that Iraq has some weapons of mass destruction. But the agency has not given President Bush a "smoking gun," according to U.S. intelligence and administration officials.

Story 13: OzText13

Drug squad detectives have asked the Police Ombudsman to investigate the taskforce that is examining allegations of widespread corruption within the squad. This coincides with the creation of a special unit within the taskforce to track the spending of at least 10 serving and former squad members. The corruption taskforce, codenamed Ceja, will check tax records and financial statements in a bid to establish if any of the suspects have accrued unexplained wealth over the past seven years. But drug squad detectives have countered with their own set of allegations, complaining to the ombudsman that the internal investigation is flawed, biased and over-zealous.

Story 14: OzText14

Queensland senator Andrew Bartlett has launched a last-minutebid to rescue the Australian Democrats from a split that threatens to destroy the party. With nominations for the party leadership to close on Wednesday night, Senator Bartlett met last night with deputy leader Aden Ridgeway to offer him a place on a unity ticket and set up a reform process to begin healing the party's wounds. Party sources said Senator Ridgeway, who turned against former leader Natasha Stott Despoja, is still expected to contest the leadership against one of her two supporters: Senator Bartlett or Brian Greig, installed as interim leader by the party's executive last Thursday.

Story 15: OzText15

Very few women have been appointed to head independent schools, thwarting efforts to show women as good leaders, according to the Victorian Independent Education Union. Although they make up two-thirds of teaching staff, women hold only one-third of principal positions, the union's general secretary, Tony Keenan, said. He believed some women were reluctant to become principals because of the long hours and the nature of the work. But in other cases they were shut out of the top position because of perceptions about their ability to lead and provide discipline.

Story 16: OzText16

The Bush administration has drawn up plans to escalate the war of words against Iraq, with new campaigns to step up pressure on Baghdad and rally world opinion behind the US drive to oust President Saddam Hussein. This week, the State Department will begin mobilising Iraqis from across North America, Europe and the Arab world, training them to appear on talk shows, write opinion articles and give speeches on reasons to end President Saddam's rule.

Story 17: OzText17

Beijing has abruptly withdrawn a new car registration system after drivers demonstrated "an unhealthy fixation" with symbols of Western military and industrial strength - such as FBI and 007. Senior officials have been infuriated by a popular demonstration of interest in American institutions such as the FBI. Particularly galling was one man's choice of TMD, which stands for Theatre Missile Defence, a US-designed missile system that is regularly vilified by Chinese propaganda channels.

Story 18: OzText18

The United Nations World Food Program estimates that up to 14 million people in seven countries - Malawi, Mozambique, Zambia, Angola, Swaziland, Lesotho and Zimbabwe - face death by starvation unless there is a massive international response. In Malawi, as many as 10,000 people may have already died. The signs of malnutrition - swollen stomachs, stick-thin arms, light-coloured hair - are everywhere.

Story 19: OzText19

In Malawi, as in other countries in the region, AIDS is making the effects of the famine much worse. The overall HIV infection rate in Malawi is 19 per cent, but in some areas up to 35 percent of people are infected. A significant proportion of the young adult population is too sick to do any productive work. Malnutrition causes people to succumb to the disease much more quickly than they do in the West, and hunger forces women into prostitution in order to feed their families, making them more vulnerable to contracting the disease. Life expectancy has been reduced to 40 years.

Story 20: OzText20

The United Nations was determined that its showpiece environment summit - the biggest

conference the world has ever witnessed - should be staged in Africa. The venue, however, could not be further removed from the grim realities of life in the rest of Africa. Johannesburg's exclusive and formerly whites-only suburb of Sandton is the wealthiest neighbourhood in the continent. Just a few kilometres from Sandton begins the sprawling Alexandra township, where nearly a million people live in squalor. Organisers of the conference, which begins today, seem determined that the two worlds should be kept as far apart as possible. Tight security surrounds Sandton's convention centre and five-star hotels, where world leaders will debate poverty, the environment and sustainable development while enjoying lavish hospitality.

Story 21: OzText21

The Iraqi capital is agog after the violent death of one of the world's most notorious terrorists, but the least of the Palestinian diplomat's worries was the disposal of Abu Nidal's body, which lay on a slab in an undisclosed Baghdad morgue. Abu Nidal's Fatah Revolutionary Council is held responsible for the death or injury of almost 1000 people in 20 countries across Europe and the Middle East in the three decades since he fell out with Yasser Arafat over what Abu Nidal saw as Arafat's willingness to accommodate Israel in the Palestinian struggle.

Story 22: OzText22

The Federal Government says changes announced today to the work for the dole scheme will benefit participants and taxpayers. Federal Employment Services Minister Mal Brough says that from July 1 those taking part in work for the dole will be able to perform extra hours to complete their mutual obligation more quickly to access training credits.

Story 23: OzText23

The biowarfare expert under scrutiny in the anthrax attacks declared, "I am not the anthrax killer," and lashed out today against Attorney General John Ashcroft for calling him a "person of interest" in the investigation. For the second time in two weeks, the scientist went before a throng of reporters outside his lawyer's office to profess his innocence and decry the attention from law enforcers that he contends has destroyed his life.

Story 24: OzText24

China said Sunday it issued new regulations controlling the export of missile technology, taking steps to ease U.S. concerns about transferring sensitive equipment to Middle East countries, particularly Iran. However, the new rules apparently do not ban outright the transfer of specific items - something Washington long has urged Beijing to do.

Story 25: OzText25

Nigerian President Olusegun Obasanjo said he will weep if a single mother sentenced to death by stoning for having a child out of wedlock is killed, but added he has faith the court system will overturn her sentence. Obasanjo's comments late Saturday appeared to confirm he would not intervene directly in the case, despite an international outcry.

Story 26: OzText26

An Islamic high court in northern Nigeria rejected an appeal today by a single mother sentenced to be stoned to death for having sex out of wedlock. Clutching her baby daughter, Amina Lawal burst into tears as the judge delivered the ruling. Lawal, 30, was first sentenced in March after giving birth to a daughter more than nine months after divorcing.

Story 27: OzText27

How did 2,300 allegedly unregistered missile warheads come to be stored on a Canadian businessman's anti-terrorism training facility in New Mexico? U.S. and Canadian officials are still trying to figure that out, but one security expert says the mystery is a "chilling" one. David Hudak, 41, was arrested in the United States more than a week ago when, according to court documents, agents searching his property found the warheads stored in crates that were marked "Charge Demolition."

Story 28: OzText28

The Saudi Interior Ministry on Sunday confirmed it is holding a 21-year-old Saudi man the FBI is seeking for alleged links to the Sept. 11 hijackers. Authorities are interrogating Saud Abdulaziz Saud al-Rasheed "and if it is proven that he was connected to terrorism, he will be referred to the sharia (Islamic) court," the official Saudi Press Agency quoted an unidentified ministry official as saying.

Story 29: OzText29

Sri Lanka's government will lift a four-year ban on Tamil Tiger rebels on Sept. 6, paving the way for peace talks with the insurgents scheduled for later that month in Thailand, a government minister said Saturday. "We will lift the ban as promised," Minister for Rehabilitation Jayalath Jayawardena told The Associated Press. The lifting of the ban is one of the key rebel conditions for resuming peace negotiations with the government after a hiatus of more than seven years.

Story 30: OzText30

A man accused of making hidden-camera footage up the skirts of women also made child pornography of the worst kind, featuring the rape of children as young as 6, police said Friday. The latest allegations suggest there's nothing humorous about voyeurs who some may perceive to be making secret videos as a joke, Staff-Insp. Gary Ellis said. "Approximately 20 per cent of voyeurs have also committed sexual assault or rape," Ellis said, reading from a recently released federal government report on criminal voyeurism.

Story 31: OzText31

Police are combing through videotapes trying to spot the gunman dressed in black who shot a 30-year-old man to death at a downtown massage parlour. The victim was hit in the stomach and upper body and died about 3 1/2 hours later in hospital. The woman was not hurt. Police urged business owners to turn over any security-camera videotapes

they might have that recorded people on the street at the time. Several such videos are now being reviewed.

Story 32: OzText32

The Federal Government did not regret its actions 12 months on from the Tampa asylum seeker crisis, Small Business Minister Joe Hockey said today. Mr Hockey said the Government was not embarrassed by the Tampa issue, which began on August 27 of last year when the captain of the Norwegian cargo ship rescued more than 400 asylum seekers from an Indonesian ferry north of Christmas Island.

Story 33: OzText33

At least three Democrats are considering splitting from the party while no-one has yet nominated to contest the leadership. Three of the "gang of four" senators who ousted Natasha Stott Despoja from the leadership are considering forming a new "progressive centre" party in the fallout from last week's turmoil. This would leave the Democrats with a rump of three or four members. West Australian Senator Andrew Murray said yesterday unless the Democrats left wing gave ground the party would split.

Story 34: OzText34

A young humpback whale remained tangled in a shark net off the Gold Coast yesterday, despite valiant efforts by marine rescuers. With its head snared by the net and an anchor rope wrapped around its tail, the stricken whale was still swimming but hopes for its survival were fading. A second rescue attempt was planned for dawn today after rescuers braved heavy seas, strong wind and driving rain to try to free the whale.

Story 35: OzText35

Prince William has told friends his mother was right all along to suspect her former protection officer of spying on her and he doesn't want any detective intruding on his own privacy. William and Prince Harry are so devastated by the treachery of Ken Wharfe, whom they looked on as a surrogate father, they are now refusing to talk to their own detectives.

Story 36: OzText36

The spectre of Osama bin Laden rose again today, urging Afghans to launch a new Jihad, or holy war, and predicting the fall of the United States, in a hand-written "letter" posted on an Islamic website. There was no hard proof that the scruffy missive was genuine, but IslamOnline.net said it had been received by their correspondent in Jalalabad, eastern Afghanistan, from an Afghan source who asked to remain anonymous. The source claimed it was the "most recent letter" from the world's most wanted man.

Story 37: OzText37

The Johannesburg Earth Summit is set to get under way with the promise that leaders will take action on the environment, debt and poverty. South African President Thabo

Mbeki, speaking at the opening ceremony, said: "Out of Johannesburg and out of Africa must emerge something that takes the world forward." But the absence of US President George W Bush was threatening to overshadow the summit.

Story 38: OzText38

Robert Mugabe strengthened his hold on the Zimbabwean government yesterday by retaining the most combative hardliner ministers in a cabinet shuffle which offered little hope of a moderation of the land seizures and other policies that have kept Zimbabwe in crisis and brought international condemnation.

Story 39: OzText39

They dress in black and disguise their identities with bandannas and sunglasses. Their logo is an image of the Southern Cross constellation, superimposed with a pair of crossed boomerangs, which resembles a swastika. The Blackshirts are former husbands aggrieved by their treatment at the hands of their ex-wives and the courts, who regard themselves as the vanguard of a "men's rights" movement in Australia and say that their actions will be remembered as marking a turning-point in history.

Story 40: OzText40

The real level of world inequality and environmental degradation may be far worse than official estimates, according to a leaked document prepared for the world's richest countries and seen by the Guardian. It includes new estimates that the world lost almost 10% of its forests in the past 10 years; that carbon dioxide emissions leading to global warming are expected to rise by 33% in rich countries and 100% in the rest of the world in the next 18 years; and that more than 30% more fresh water will be needed by 2020.

Story 41: OzText41

Researchers conducting the most elaborate wild goose chase in history are digesting the news that a bird they have tracked for over 4,500 miles is about to be cooked. Kerry, an Irish light-bellied Brent goose, was one of six birds tagged in Northern Ireland in May by researchers monitoring the species' remarkable migration. Last week, however, he was found dead in an Inuit hunter's freezer in Canada, still wearing his 3,000 satelite tracking device. Kerry was discovered by researchers on the remote Cornwallis Island. They picked up the signal and decided to try to find him.

Story 42: OzText42

Russia defended itself against U.S. criticism of its economic ties with countries like Iraq, saying attempts to mix business and ideology were misguided. "Mixing ideology with economic ties, which was characteristic of the Cold War that Russia and the United States worked to end, is a thing of the past," Russian Foreign Ministry spokesman Boris Malakhov said Saturday, reacting to U.S. Defense Secretary Donald Rumsfeld's statement that Moscow's economic relationships with such countries sends a negative signal.

Story 43: OzText43

Pope John Paul II urged delegates at a major U.N. summit on sustainable growth on Sunday to pursue development that protects the environment and social justice. In comments to tourists and the faithful at his summer residence southeast of Rome, the pope said God had put humans on Earth to be his administrators of the land, "to cultivate it and take care of it." "In a world ever more interdependent, peace, justice and the safekeeping of creation cannot but be the fruit of a joint commitment of all in pursuing the common good," John Paul said.

Story 44: OzText44

The Russian defense minister said residents shouldn't feel threatened by the growing number of Chinese workers seeking employment in the country's sparsely populated Far Eastern and Siberian regions. There are no exact figures for the number of Chinese working in Russia, but estimates range from 200,000 to as many as 5 million. Most are in the Russian Far East, where they arrive with legitimate work visas to do seasonal work on Russia's low-tech, labor-intensive farms.

Story 45: OzText45

Australian spies listened to conversations between Norway's ambassador and its foreign office during the Tampa crisis, a soon to be published book will reveal. Phone calls were tapped by the Defence Signals Directorate when Norwegian ambassador Ove Thorsheim visited the freighter during the stand-off. A book, Tampa, to be published in Norway in October, recounts the events which triggered Australia's Pacific Solution and transformed Tampa Captain Arne Rinnan into a homeland hero.

Story 46: OzText46

Batasuna, a political party that campaigns for an independent Basque state, faces a double blow today: the Spanish parliament is expected to vote overwhelmingly in favour of banning the radical group, while a senior investigative judge is poised to suspend Batasuna's activities on the grounds that they benefit Eta, the outlawed Basque separatist group.

Story 47: OzText47

The river Elbe surged to an all-time record high Friday, flooding more districts of the historic city of Dresden as authorities scrambled to evacuate tens of thousands of residents in the worst flooding to hit central Europe in memory. In the Czech Republic, authorities were counting the cost of the massive flooding as people returned to the homes and the Vlava river receded, revealing the full extent of the damage to lives and landmarks.

Story 48: OzText48

The European Parliament is spoiling for a fight with Israel. It has voted to review the EU's diplomatic links with the Jewish state, to impose an arms embargo and to threaten wider trade sanctions. Many MEPs want to go further and dispatch a European military force to the region in order to "protect the Palestinian people".

Story 49: OzText49

Australia's Commonwealth Bank on Wednesday said it plans to cut about 1,000 jobs even as it reported its profit rose 11 percent last fiscal year. Workers reacted angrily to the planned cuts, which Australia's second largest bank said were designed to control costs. The cuts will take effect this financial year. The bank reported net profit of 2.66 billion Australian dollars ($1.4 billion) in the year to June 30, up from 2.4 billion Australian dollars in the previous year.

Story 50: OzText50

Labor needed to distinguish itself from the Government on the issue of asylum seekers, Greens leader Bob Brown has said. His Senate colleague Kerry Nettle intends to move a motion today - on the first anniversary of the Tampa crisis - condemning the Government over its refugee policy and calling for an end to mandatory detention. "We Greens want to bring the Government to book over its serial breach of international obligations as far as asylum seekers in this country are concerned," Senator Brown said today.

A.4 Google News Experiment Texts

Story 1: 1_news.xinhuanet.com.HOSTAGES.txt, 10 October 2004

1. Pakistani minister links captors of Chinese engineers with al-Qaeda.
2. ISLAMABAD, Oct. 10 (Xinhuanet)
3. Pakistan Interior Minister Aftab Ahmed Khan Sherpao Sunday said that al-Qaeda has links with the kidnapping of two Chinese engineers, who were kidnapped in the country's tribal area on Saturday.
4. The Chinese engineers, along with a Pakistani guard, were abducted near Jandola in Pakistan's South Waziristan tribal area, bordering Afghanistan, by five gunmen.
5. **The Chinese hostages, Wang Ende and Wang Peng, were working on a water dam and a canal in the tribal area for China National Water Resources and Hydropower Engineering Group Corporation. "**
6. There is definitely a link between the captors and al-Qaeda.
7. Two of the captors are foreigners and they may belong to Central Asia.
8. There is information that two belong to Kandahar in Afghanistan, Sherpao told local Geo television.
9. He said two of the kidnappers have covered faces while the another two are Pakistanis. "
10. I am sure that the Chinese engineers, their guard and driver are safe and are in the control of a local tribe Jalal Khel tribe, he said adding that the kidnappers are also in the custody of the tribe.
11. He said a 21-member jirga or council is holding talks with the tribe.
12. **Member of the National Assembly from the tribal region Maulana Mirajuddin is also member of the jirga.**
13. Sherpao said the captors have links with those who are the targets of an operation in the area of Mahsud tribe.
14. He said that demands of the captors would be clear after the jirga end negotiations. "
15. **The government will look into the demands whether or not they are acceptable the minister said.**
16. Pakistani army has been engaged in operation to flush out hundreds of Al-Qaeda linked militants from South Waziristan since last year.

Story 2: 3_www.news.com.au.HOSTAGES.txt, 10 October 2004

1. Kidnappers threaten to kill Chinese.
2. MILITANTS holding two Chinese engineers in Pakistan's rugged tribal area have threatened to kill them unless al-Qaeda detainees are released, a security official said today. "
3. **There is only one demand.**
4. **They have asked for the release of all arrested foreign militants, the intelligence official based in the northwest city of Peshawar told AFP on condition of anonymity. "**
5. They have given one day to the government to consider their demand, otherwise they said they will kill the hostages.
6. The four captors three foreign militants and a local tribesman issued the demands yesterday
7. **China's embassy in Islamabad was unaware of the threat and demand and had been told the engineers were safe and in the custody of a local tribe. "**

8. According to our information at this moment our two Chinese engineers are still safe but they have yet to be released, spokesman Zhang Yiming told AFP. "
9. We are informed that the Chinese engineers and the kidnappers are separately under the custody of local tribal elders, and that they are safe this moment.
10. North West Frontier provincial home secretary Abdul Karim Qasuria said the local Mahsud tribe was negotiating with the militants holding the engineers. "
11. They are hopeful of ending the crisis peacefully he said. "
12. I believe they are militants, he said of the captors, denying any local people were holding them.
13. The engineers were kidnapped yesterday morning as they travelled from the western town of Dera Ismail Khan to the Gomalzam dam project in South Waziristan, the remotest and most dangerous of Pakistan's north-west tribal districts.
14. Hundreds of al-Qaeda militants have taken refuge and set up training camps in South Waziristan, and have been battling thousands of Pakistani army troops trying to flush them out

Story 3: 1_news.xinhuanet.com.SOMALIA.txt, 10 October 2004

1. 6 candidates go to 2nd round of Somali presidential elections.
2. NAIROBI, Oct. 10 (Xinhuanet)
3. Six of the 26 candidates will vie for the second round of Somalia's presidential elections in Kenya's capital Nairobi on Sunday, after a first ballot failed to give any candidate the two-thirds majority required for outright victory, a Xinhua correspondent at the voting site said.
4. **Unofficial results show that Abdullahi Yusuf Ahmed, 71, leader of the autonomous Somali region of Puntland, gained 80 votes in the first round of the elections, leaving the other 25 contestantsfar behind.**
5. Somali lawmakers started voting for president of the Horn of Africa country on Sunday.
6. The new president is expected to appoint a prime minister to form a government, mandated to lead Somalia through a five-year transitional period.
7. **A total of 28 contestants including a woman were competing for the presidency in what is widely seen as the culmination of a tortuous two-year peace process in Kenya, mediated by the regional Inter-Governmental Authority on Development (IGAD)**
8. **However two presidential aspirants withdrew from the contest at the last minute, leaving 26 candidates to take an oath binding them to abide by whatever results from the voting.**
9. According to election organizers, in the likely event of no candidate obtaining more than two-thirds of the ballot, the top six runners will go to a second round.
10. If none of them gains a two-thirds majority, the two leading candidates will face off in a third round on a simple majority basis.
11. Since the breakdown of the Somali central government in 1991, conflict and famine have killed hundreds of thousands of people, plunging the country into anarchy.
12. Under the auspices of IGAD, which groups Djibouti, Ethiopia, Eritrea, Kenya, Sudan, Uganda and Somalia, Somali National Reconciliation Conference began in October 2002 in Kenya.

13. The presidential elections come after the election of the speaker of the transitional federal parliament of Somalia on September 15 and his two deputies on September 22 and the inauguration of the transitional parliament on August 29.

Story 4: 2_www.news24.com.SOMALIA.txt, 10 October 2004

1. Nairobi - Somali members of parliament on Sunday began the process of electing a new president for their anarchic Horn of Africa state at a ceremony in the Kenyan capital.
2. The election process, held in Kenya because of inadequate security in Somalia itself, began with a speech from the speaker of Somalia 's newly formed national assembly at a packed sports stadium on the outskirts of Nairobi.
3. Twenty-eight candidates are running for president.
4. They include Abdulkassim Salat Hassan, who was elected president of Somalia's Transitional National Government in 2000 by a previous legislature, but who never managed to exert authority far beyond a few pockets of the battle-scarred capital.
5. **The incoming president is immediately to appoint a prime minister, whose government will be the fourth of its kind to be set up outside Somalia by peace conferences since 1991, when the country's last central government, led by dictator Mohamed Siad Barre, was toppled.**
6. The election is the culmination of two difficult years of talks in Kenya, which differed from previous negotiations in that they included most of the warlords who have battled for control of parts of Somalia since Barre's fall.
7. **Since under power-sharing deals the jobs of president and prime minister must be distributed between two of Somalia's four largest clans, the Hawiye and Darod, there are clear frontrunners in the race for the head of state.**
8. **As well as Salat (Hawiye) they are Abdullahi Yusuf Ahmed (Darod) the current president of Puntland, a semi-autonomous state in the northeast and former finance minister Abdullahi Mohamud Addow (Hawiye)**
9. The 275 MPs will cast their votes by secret ballot.
10. In the likely event of no candidate winning more than two-thirds of the ballot, the top six runners will go to a second round.
11. If none secures a two-thirds majority, the two leading candidates will face off in a third round on a simple majority basis.

Story 5: 1_washingtontimes.com.HAITI.txt, 10 October 2004

1. U.N. peacekeeper shot in Haitian melee.
2. Port-au-Prince, Haiti, Oct. 10 (UPI)
3. A U.N. peacekeeper was wounded in a gun battle with Haitian gangs loyal to former President Jean-Bertrand Aristide, Haitian television reported Sunday.
4. **The soldier, a member of the Brazilian troops leading the peacekeeping mission, was shot in the hostile Bel Air neighborhood of Port-au-Prince, where Aristide supporters have often gathered to call for his return to power.**
5. The former Haitian leader left the country in February amid increasing pressure from armed rebel groups calling for his resignation.

6. **He currently has asylum in South Africa, although he says publicly he intends to return one day.**
7. Saturday's shooting marked the first time a soldier has been wounded during the four-month U.N. mission.
8. According to a spokesman for the Brazilian contingent in Haiti, the soldier was shot in his ankle during a sweep, backed by Haitian police, of the slum neighborhood.
9. **He is expected to make a full recovery.**

Story 6: 2_seattletimes.nwsource.com.HAITI.txt, 10 October 2004

1. Gunmen in Haiti wound U.N. peacekeeper.
2. PORT-AU-PRINCE, Haiti.
3. A gunbattle broke out between U.N. peacekeepers and supporters of ousted Haitian President Jean-Bertrand Aristide yesterday wounding a peacekeeper for the first time in the force 's 4-month- old mission.
4. **Heavy gunfire erupted in the capital when Brazilian troops backed by Haitian police rolled through the volatile slum of Bel Air, where armed men were demanding the return of Aristide from exile, said U.N. spokesman Toussaint Kongo-Doudou.**
5. Peacekeepers " came under heavy fire and they returned fire, Kongo-Doudou said, adding that it appeared that some of the gunmen were wounded.
6. The Brazilian soldier was wounded in the foot, the first casualty among some 3,000 peacekeepers.
7. **Kongo-Doudou said troops and police arrested more than 60 people suspected of attacking them.**

Story 7: 3_news.xinhuanet.com.HAITI.txt, 10 October 2004

1. UN soldier injured in clash with Aristide loyalists.
2. HAVANA, Oct. 9 (Xinhuanet)
3. A UN peacekeeper and a Haitian policeman were injured Saturday in a clash with loyalists of former Haitian President Jean-Bertrand Aristide in Port-au-Prince.
4. According to reports from the capital of Haiti, the victims were attacked by armed groups during an operation in the Bel Air neighborhood, a stronghold of pro-Aristide forces.
5. **The police have detained some mobsters and captured many weapons, said Haiti's General Police Director Leon Charles, addingsome militants are hiding in a Catholic church of Bel Air.**
6. Meanwhile UN Mission for the Stabilization of Haiti (MINUSTAH) spokesman Toussaint-Kongo Doudou indicated that it is impossible to determine the number of casualties among the pro-Aristide forces right now as the violence is continuing.
7. On Friday, four loyalists of Aristide died in a skirmish with the police.
8. **Officials said three police officers were beheaded by Aristide supporters last week as part of a campaign called Operation Baghdad.**
9. Dozens have been wounded in shootings.

10. At least 26 people have been killed in violence that erupted as Aristide supporters stepped up protests on Sept. 30, demanding their leader's return from exile in South Africa and an end to " the invasion " of foreign troops, referring to US Marines who arrived as Aristide left in February and UN peacekeepers who took over in June.
11. **Relief organizations feared the escalation in violence in Port-au-Prince would paralyze the distribution of food to tens of thousands of people after the devastating tropical storm Jeanne.**
12. **At least 2,500 Haitians were killed or reported missing in the disaster last month.**

Story 8: 5_www.cbc.ca.HAITI.txt, 10 October 2004

1. Brazilian soldier wounded in gunbattle with Aristide loyalists in Haiti.
2. PORT-AU-PRINCE, Haiti.
3. Gunbattles broke out Saturday as UN troops entered a volatile slum dominated by supporters of ousted Haitian president Jean-Bertrand Aristide, wounding at least one Brazilian soldier.
4. **In northern Haiti, protesters crowded outside a mass for flood victims, saying the government hasn't done enough to help survivors.**
5. **Gunfire erupted as Brazilian troops backed by Haitian police rolled through the volatile slum Bel Air, where Haitians demanding the return of Aristide from exile had blocked streets with car chassis and debris in recent days, UN spokesman Toussaint Kongo-Doudou said.**
6. He said one soldier was shot and wounded and some of the gunmen appeared to have been wounded.
7. Police said one officer also was wounded in the leg by gunfire in the nearby Sans Fil neighbourhood.
8. **Police detained at least 34 men in the area, holding some to the ground at gunpoint before loading them into a police truck.**
9. In the Cite Soleil slum, residents also reported gunfire.

Story 9: 3_pakistantimes.net.DERRIDA.txt, 10 October 2004

1. PARIS (France) French philosopher Jacques Derrida has died of cancer at the age of 74 on Saturday, said sources.
2. Jacques Derrida, the founder of the school of deconstructionism, has died of cancer at the age of 74, France Info radio said on Saturday.
3. It said Algerian-born Derrida had died on Friday of cancer of the pancreas.
4. **Derrida, who divided his time between France and the United States, argued that the traditional way we read texts makes a number of false assumptions and that they have multiple meanings which even their author may not have understood.**
5. **His thinking gave rise to the school of deconstruction, a method of analysis that has been applied to literature, linguistics, philosophy, law and architecture.**
6. It is heralded as showing the multiple layers of meaning at work in language, but was described by critics as nihilistic "

7. In him, France gave the world one of the greatest contemporary philosophers, one of the major figures in the intellectual life of our time, French President Jacques Chirac said in a statement after learning of his death.
8. **Through his work, he sought to find the free movement which lies at the root of all thinking.**

Story 10: 4_observer.guardian.co.uk.DERRIDA.txt, 10 October 2004

1. World's greatest philosopher dies.
2. Jacques Derrida, the French thinker once described as the most influential philosopher in the world, has died at the age of 74 in a Paris hospital.
3. The controversial theorist was diagnosed with aggressive pancreatic cancer last year.
4. **Derrida, whose death was announced by the office of the French President Jacques Chirac, has been credited with the invention of three philosophical concepts which dominated late 20th century thinking : 'postmodernism"', 'poststructuralism"' and 'deconstruction "', though in later years he showed growing irritation as the words passed into daily use.**
5. In a recent interview, he said the word 'deconstruction "' had even penetrated a description of a rabbit stew which he had seen in a newspaper.
6. **'Deconstructed rabbit!**
7. I saw it in an article in the New York Times,"' he said.
8. Derrida grew up in El-Biar, Algeria, moving to France at the age of 19.
9. From 1952 he studied at Paris's Ecole Normale Superieur under two French philosophical greats, Michel Foucault and Louis Althusser.
10. Although he faced deep suspicion from the British philosophical tradition for his grand theorising - 20 philosophers objected to his honorary degree granted by Cambridge - he remained a towering figure in world philosophy.
11. **He died with one great disappointment : he would have liked to follow in the footsteps of his illustrious predeccesor, the writer Albert Camus, who played in goal for Algeria's football team.**
12. 'I wanted to be a professional soccer player, but I had to give it up because I was not good enough"' he once said.

Story 11: 4_www.ananova.com.BIGLEY.txt, 10 October 2004

1. Tape shows final plea of hostage.
2. A videotape of the beheading of British hostage Kenneth Bigley has appeared on an Islamist website, showing the civil engineer pleading that he wanted to live a simple life moments before he was decapitated.
3. The nearly five-minute-long tape appeared two days after Mr Bigley's family said it had proof that the 62-year- old civil engineer from Liverpool was dead
4. **The body has not been found.**
5. The tape showed Mr Bigley, dressed in an orange prison-style jump suit and seated in front of seven armed hooded men.
6. **Behind them was a banner of the feared Tawhid and Jihad group, the extremist organisation which has claimed responsibility for numerous suicide attacks and beheadings of Westerners.**

7. Bigley made a brief statement, saying I am not a difficult person.
8. I am a simple man who wants to live a simple life.
9. He then told British Prime Minister Tony Blair that more than ever I need your help.
10. Here I am again, Mr Blair, Mr Bigley added.
11. Very very close to the end of my life, you do not appear to have done anything at all to help me.
12. He said his captors patience is wearing thin and they are very very serious people.
13. Please, please give them what they require – freedom of their women from the Abu Ghraib prison, Mr Bigley said calmly.
14. I beg you British people, more then ever I need your help, more than ever I need your voices, to go out in the street and to demand a better life for the females and the women who are imprisoned in the Abu Ghraib prison near Baghdad.
15. All I can tell you now is that I have very short time left.

Story 12: 5_www.abc.net.au.BIGLEY.txt, 10 October 2004

1. Video shows Bigley's last appeal.
2. A video tape posted on the Internet shows British hostage Kenneth Bigley making a final appeal to the British Government to meet his captors' demands before he was beheaded by a militant group in Iraq.
3. Here I am again Mr (Tony) Blair and your Government, very very close to the end of my life, the tape showed the 62-year- old engineer saying before masked men severed his head with a knife.
4. I 'm not a difficult person.
5. I am a simple man who just wants to live a simple life with his family.
6. These people, their patience is wearing very very thin and they are very serious people.
7. Please, please give them what they require, the freedom of the women in Abu Ghraib prison.
8. If you do this the problem is solved, said Mr Bigley, who was sitting on the floor with militants standing behind him.
9. Insurgent sources in Iraq had said the Briton escaped briefly from his captors shortly before they killed him in a town south-west of Baghdad.
10. **He was kidnapped on September 16, along with two Americans who were beheaded soon after by the Tawhid and Jihad group led by Al Qaeda ally Abu Musab al-Zarqawi, which demanded the release of women prisoners from United States jails in Iraq.**
11. **To the British people, more than ever I need your help, more than ever I need your voices, to go out into the streets and demand a better life for the females and women who are in prison in Abu Ghraib, said Mr Bigley, who was wearing an orange jump suit like those worn by detainees in US prisons.**
12. **The four-minute tape, posted on Web sites often used by Islamists, showed militants holding up the Briton 's severed head and then placing it on top of the corpse.**

Story 13: 2_news.xinhuanet.com.SINAI.txt, 10 October 2004

1. Blast rescue work nearly ends, Bedouin tribemen detained for questioning.

2. CAIRO, Oct. 10 (Xinhuanetnet)
3. The search-rescue work of the Taba blast, which killed at least 34 people and wounded 124, at Red Sea resorts in south Sinai of Egypt on Thursday night, was nearly completed, and some 15 Bedouin tribemen were detained by police on suspicion of involvement in the attack.
4. Israeli search and rescue teams said they expected to complete their search work later Sunday at the blast site of the hotel as no more bodies were found under the rubble.
5. A deadly blast Thursday night rocked Hilton Taba hotel in Egypt's Sinai Peninsular, which was popular with Israeli tourists.
6. Shortly after that, two other blasts rocked two holiday camps in Ras al-Sultan and Nuwaiba, not far away from Taba.
7. Thousands of Israeli tourists were in the Red Sea resorts of the Sinai Peninsula over the Jewish holiday of Sukkot despite travelling warnings given by Israeli security officials last month, saying they received information of attacks against Israelis in the resort area.
8. Meanwhile the investigation into the terror attacks was underway.
9. On Saturday, investigators lifted fingerprints, swabbed dust and collected tissue from the sites of three car bombings and some 15 Bedouin tribesmen were detained for allegedly providing the explosives used in the attacks.
10. The tribemen were believed to have helped smuggle the explosive used in the attacks, security sources said on Sunday.
11. Israel has blamed al-Qaida for the attacks.
12. But the Egyptian security sources said it was too early to jump a conclusion, saying at the moment they are concentrating their efforts on identifying the vehicles and the source of the explosives.
13. The United States is still reserved regarding who has carried out these operations, with US State Department Spokesman Richard Boucher saying that Washington could not define who was involved in the attacks.
14. Egyptian officials said 34 people were killed, including nine Egyptians, at least five Israelis, in the most serious terrorist attack in Egypt since 1997 when 58 foreign tourists were massacred by Islamic militants in Luxor in southern Egypt.
15. **On Sunday, the Israeli cabinet discussed a proposal of counter-terror unit on imposing economic sanctions on Israelis who travel to destinations which were classified in travel advisories as dangerous the Ha'aretz daily reported Sunday.**
16. **The proposal, which was put forward following the three blasts, would include deprival of rights to compensation guaranteed to victims of terror attacks, the paper said.**
17. **Sanctions of the kind are already being practiced by the United States and Britain.**
18. **At Sunday's meeting, Israeli Health Minister Danny Naveh said National Security Advisor Giora Eiland is considering the matter and is to determine clear guidelines for situations in which the Shin Bet warns Israelis against traveling to Sinai or other destinations.**
19. The paper said earlier there are currently thousands of Israelis staying in the Peninsula.

Story 14: 4_www.news24.com.SINAI.txt, 10 October 2004

1. Taba - Israeli rescue workers wound down their search on Sunday for victims of a luxury hotel bombing cutting through debris with their Egyptian counterparts in a dusty hole that was a basement shopping area.
2. Later on Sunday, the Israelis rescue and recovery teams expected to finish their work and return to Israel.
3. At least 34 people were killed in the Taba blast and two smaller car bomb explosions Thursday night at Sinai tourist resorts. "
4. We are at the last stage of recovery, said Gideon Baron, head of the Israeli rescue and recovery effort.
5. No bodies had been found since Saturday night, but Baron would not say if more were believed buried
6. While the Egyptian-Israeli rescue and recovery effort wound down, the investigation was moving ahead.
7. On Saturday, investigators lifted fingerprints, swabbed dust and collected tissue from the sites of three car bombings and detained dozens of Bedouin tribesmen, including quarry workers who could have provided the explosives.
8. Egyptian investigators said they suspect a group of eight to 10 terrorists targeting Israelis carried out the Thursday night attacks, possibly slipping in from Saudi Arabia or Jordan on speed boats.
9. Al-Qaeda blamed Israel has blamed al-Qaeda for the attacks.
10. The Egyptian investigators are leaning toward an al-Qaeda connection as well saying a local sleeper cell may have been awakened to carry out the attacks, Egypt's first terrorist strike in seven years.
11. **Speaking on condition of anonymity, they said such a group would almost certainly be linked to Ayman al-Zawahri, who led the Egyptian Islamic Jihad before merging his group with al-Qaeda in 1998.**
12. The Egypt-born Zawahri is now bin Laden's top deputy.
13. Three car bombs exploded Thursday night, one at the Taba Hilton just south of the Egypt-Israel border and two at a town of beach bungalows, Ras Shitan, 55km to the south on the Red Sea coast.
14. **Two Egyptian security officials, speaking on condition of anonymity, said at least one woman saw two men in the car believed used in the Taba attack before it exploded and was able to provide detailed descriptions. "**
15. **It was dark and she could not give 100 details, but the information she gave is good and provided some clues about how it all happened, one official said.**
16. One of the officials said investigators were leaning toward the foreigners' scenario because of the sophistication of the attacks, describing a home-grown scenario as a 10-15 % possibility.
17. The other said it could have been planned abroad and had some Egyptian involvement.
18. Israeli officials said they believed al-Qaeda was behind the attack. "
19. The type, the planning, the scope, the simultaneous attacks in a number of places, all this points to al-Qaeda Israeli Foreign Minister Silvan Shalon told Channel 2.
20. US officials share the suspicion.

Story 15: Arsenic1.txt, via Google News, 3 May 2003

1. Maine Police Link Dead Man to Arsenic Case Sat May 3, 2003 01 : 32 PM CARIBOU, Maine (Reuters)— Maine police on Saturday linked a man who died of a possibly self-inflicted gunshot wound to the arsenic-tainted coffee poisoning at a local church that killed one parishioner and sickened at least 15 others.
2. State Police Col. Michael Sperry told reporters that Daniel Bondeson, 53, who died on Friday evening at a hospital in Caribou, was involved in the poisoning at the church that has rocked New Sweden, Maine, a town of about 600 people near the Canadian border. "
3. We have linked the shooting to the death at the church, Sperry said, explaining that police had found key information at Bondeson's farm house where he was found shot.
4. Bondeson died after emergency surgery.
5. Sperry however declined to describe the information that police had found, and declined to say whether police had discovered a suicide note or arsenic.
6. A search of Bondeson's Woodland, Maine home was expected to take several days. "
7. This was reported as a self-inflicted gunshot wound to us, Sperry said, adding however that police won't say for sure what happened until the body is autopsied on Monday.
8. The mystery began last Sunday when parishioners at the Gustaf Adolph Lutheran Church in New Sweden suddenly fell ill after services.
9. One man, 78-year- old Walter Morrill, died after drinking the coffee and several others were hospitalized.
10. Bondeson, a bachelor, was a member of the church but did not attend last week's sermon.
11. However his brother and sister were there but did not drink anything, parishioners said.
12. Police earlier this week said the concentration of arsenic led them to believe it had been put into the coffee deliberately
13. They called in the Federal Bureau of Investigation to gather fingerprints and DNA samples from as many as 50 parishioners to try and find the killer.
14. **Arsenic can kill quickly if consumed in large quantities, although small long-term exposure can lead to a much slower death.**
15. **It can give a strong bitter taste to food or beverages it contaminates.**

Story 16: Arsenic2.txt, via Google News, 3 May 2003

1. Shooting may be tied to poisoning
2. 5/3 / 2003 1 : 06 PM.
3. NEW SWEDEN, Maine (UPI)— Investigators in the small northern Maine town of New Sweden searched Saturday for possible links between a fatal shooting and a church arsenic poisoning case.
4. Daniel Bondeson, 53, died Friday evening, shortly after he was found shot at his home in Woodland, adjacent to New Sweden.
5. Police said Bondeson was not a suspect in the church poisoning but searched his home for any connection to the laced church coffee that killed one elderly man and sickened at least 15 other people.
6. Two of the victims were still in critical condition.
7. " We won't make a determination until the autopsy (Bondeson's body) Monday, Col. Michael Sperry, chief of the Maine State Police, said Saturday outside the Caribou courthouse.

8. " This is an open investigation, and we are still looking at who is involved, he said.
9. Investigators are treating the arsenic case as a homicide.
10. " The FBI is still very much involved adding the arsenic poisoning has had " a huge impact on this small community, about 20 miles south of the Canadian border.
11. The poisoning occurred Sunday after services at the Gustaf Adolph Lutheran Church in New Sweden, a community of about 620 in the potato-farming region in rural northern Maine.
12. Among the two dozen parishioners who drank the coffee was Walter Reid Morrill, 78, who died Monday.
13. **Morrill was a member who lived next to the church and served as caretaker and head usher.**
14. The remaining coffee in the percolator contained " high levels " of arsenic, said Stephen McCausland, spokesman for the Maine Department of Public Safety.
15. He said tests confirmed the arsenic was not in the unbrewed coffee, in the tap water or in the sugar.
16. Word that someone apparently had deliberately put arsenic in the coffee frightened area residents.
17. **" This is a small community where everyone knows each other McCausland told United Press International.**
18. We don't know whether (the perpetrator) was among them or someone from outside, but the focus of our investigation now is to find who is responsible for introducing the arsenic and why

Appendix B

WordNet Specificity Experiment Materials

In the following pages, each lettered line defines a CONCEPT, giving the words used to express that concept, a short definition of the concept and occasionally an example of the words in use in a sentence.
Please order these sets of concepts from most general to least general
If you think 2 or more concepts are equally general, rank them at the same level
If you think it is not possible to rank the concepts, mark the set with a star.

Set I Put in order

a silicone, silicone polymer – (any of a large class of siloxanes that are unusually stable over a wide range of temperatures; used in lubricants and adhesives and coatings and synthetic rubber and electrical insulation)
b soul, psyche – (the immaterial part of a person; the actuating cause of an individual life)
c metal filing – (a fragment of metal rubbed off by the use of a file)

1 Most general:
2 Middle degree of generality:
3 Least general:

Set II Put in order

a trout – (any of various game and food fishes of cool fresh waters mostly smaller than typical salmons)
b table knife – (a knife used for eating at dining table)
c blues – (a type of folk song that originated among Black Americans at the beginning of the 20th century; has a melancholy sound from repeated use of blue notes)

1 Most general:
2 Middle degree of generality:
3 Least general:

Set III Put in order

a vitamin B2, vitamin G, riboflavin, lactoflavin, ovoflavin, hepatoflavin – (a B vitamin that prevents skin lesions and weight loss)
b scapegoat, whipping boy – (someone punished for the errors of others)
c vent, volcano – (a fissure in the earth's crust (or in the surface of some other planet) through which molten lava and gases erupt)

1 Most general:
2 Middle degree of generality:
3 Least general:

Set IV Put in order

a English person – (a native or inhabitant of England)
b major suit – ((in bridge) a suit of superior scoring value, either spades or hearts)
c gymnosperm – (plants of the class Gymnospermae having seeds not enclosed in an ovary)

1 Most general:
2 Middle degree of generality:
3 Least general:

Set V Put in order

a digit, dactyl – (a finger or toe in human beings or corresponding part in other vertebrates)
b magpie – (long-tailed black-and-white bird that utters a chattering call)
c provisions, provender, viands, victuals – (a stock or supply of foods)

1 Most general:
2 Middle degree of generality:
3 Least general:

Set VI Put in order

a ribbon, thread – (any long object resembling a thin line; "a mere ribbon of land"; "the lighted ribbon of traffic"; "from the air the road was a gray thread"; "a thread of smoke climbed upward")
b allergen – (any substance that can cause an allergy)
c stretch – (a large and unbroken expanse or distance; "a stretch of highway"; "a stretch of clear water")

1 Most general:
2 Middle degree of generality:
3 Least general:

Set VII Put in order

a filing – (a fragment rubbed off by the use of a file)
b foodstuff, food product – (a substance that can be used or prepared for use as food)
c bit, chip, flake, fleck, scrap – (a small fragment of something broken off from the whole; "a bit of rock caught him in the eye")

1 Most general:
2 Middle degree of generality:
3 Least general:

Set VIII Put in order

a lab, laboratory, research lab, research laboratory, science lab, science laboratory – (a workplace for the conduct of scientific research)
b fishery, piscary – (a workplace where fish are caught and processed and sold)

c farm – (workplace consisting of farm buildings and cultivated land as a unit; "it takes several people to work the farm")

1 Most general:
2 Middle degree of generality:
3 Least general:

Set IX Put in order

a sea boat – (a boat that is seaworthy; that is adapted to the open seas)
b wall – (anything that suggests a wall in structure or effect; "a wall of water"; "a wall of smoke"; "a wall of prejudice")
c printmaker, graphic artist – (an artist who designs and makes prints)

1 Most general:
2 Middle degree of generality:
3 Least general:

Set X Put in order

a axis, axis of rotation – (the center around which something rotates)
b river boat – (a boat used on rivers or to ply a river)
c remains – (any object that is left unused or still extant; "I threw out the remains of my dinner")

1 Most general:
2 Middle degree of generality:
3 Least general:

Set XI Put in order

a marginalization – (the social process of becoming or being made marginal (especially as a group within the larger society); "the marginalization of the underclass")
b boric acid, boracic acid – (any of various acids containing boron and oxygen)
c volume, loudness, intensity – (the magnitude of sound (usually in a specified direction); "the kids played their music at full volume")

1 Most general:
2 Middle degree of generality:
3 Least general:

Set XII Put in order

a rational motive – (a motive that can be defended by reasoning or logical argument)
b disapproval – (the act of disapproving or condemning)
c harmony, concord, concordance – (agreement of opinions)

1 Most general:
2 Middle degree of generality:
3 Least general:

Set XIII Put in order

a average, norm – (a statistic describing the location of a distribution: "it set the norm for American homes")
b capture, gaining control, seizure – (the act of forcibly dispossessing an owner of property)
c animal skin – (the outer covering of an animal)

1 Most general:
2 Middle degree of generality:
3 Least general:

Set XIV Put in order

a immune response, immune reaction, immunologic response – (a bodily defense reaction that recognizes an invading substance (an antigen: such as a virus or fungus or bacteria or transplanted organ) and produces antibodies specific against that antigen)
b cut, cutting – (the act of cutting something into parts; "his cuts were skillful"; "his cutting of the cake made a terrible mess")
c praise, congratulations, kudos – (an expression of approval and commendation; "he always appreciated praise for his work")

1 Most general:
2 Middle degree of generality:
3 Least general:

Set XV Put in order

a psychiatry, psychopathology, psychological medicine – (the branch of medicine dealing with the diagnosis and treatment of mental disorders)
b vowel – (a letter of the alphabet standing for a spoken vowel)
c agenda, agendum, order of business – (a list of matters to be taken up (as at a meeting))

1 Most general:
2 Middle degree of generality:
3 Least general:

Appendix C

StoryLines Materials and Source Code

C.1 LT POS tags

Taken from LT POS website: **http://www.ltg.ed.ac.uk/software/penn.html**. LT POS can be used with the Penn Treebank tagset. The most important tags are listed below. For the complete set, see (Marcus *et al.* 1993).

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	there is
FW	foreign word	d'hoevre
IN	preposition/subordinating conjunction	in, of, like
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	predeterminer	both the boys
POS	possessive ending	friend's
PRP	personal pronoun	I, he, it
PRP$	possessive pronoun	my, his
RB	adverb	however, usually
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give up
TO	to	to go, to him
UH	interjection	uhhuhhuhh
VB	verb, base form	take
VBD	verb, past tense	took
VBG	verb, gerund/present participle	taking
VBN	verb, past participle	taken
VBP	verb, sing. present, non-3d	take
VBZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when

C.2 Step 2 Term Extraction source code

```
#!usr/bin/perl -w

use File::Basename;
use IPC::Open2;
use strict;

my ($in, $indir, $outdir, $ext, @allfiles, $infile, $inputArray,
$basename, $nfile_bare, $nfile, $vfile, $afile, $advfile, $txtfile);

if($#ARGV!=1){
    die "Usage: perl $0 <input directory> <output directory>\n";
}
else {
    $in=$ARGV[0];
    $outdir=$ARGV[1];
}

if(-d $in){
    $indir=$in;
    opendir(DIR, "$indir") or die "couldn't open $indir: $!";
    @allfiles=grep !/^\.\.?\z/, readdir DIR;
}
else {
    ($allfiles[0], $indir, $ext) = &fileparse($in);
    if(defined $ext){
        $allfiles[0]=$allfiles[0].$ext;
    }
 }

for my $j (0 .. $#allfiles){
    $infile=join('/', $indir, $allfiles[$j]);
    $inputArray=readInputFile($infile);
    #### OPEN OUTPUT FILES ####
    $basename=&basename($infile);
    $txtfile=join "", $outdir, $basename, ".txt";
    &printIndexedSentencesText($inputArray, $txtfile);
    $nfile_bare=join "", $outdir, $basename, ".np_bare";
    $nfile=join "", $outdir, $basename, ".nps";
    $vfile=join "", $outdir, $basename, ".vps";
    $afile=join "", $outdir, $basename, ".adj";
```

```
$advfile=join "", $outdir, $basename, ".adv";
open(NPS_BARE, ">$nfile_bare")
              or die "Can't open np file: $nfile_bare\n";
open(NPS, ">$nfile") or die "Can't open np file: $nfile\n";
open(VPS, ">$vfile") or die "Can't open vp file: $vfile\n";
open(ADJ, ">$afile") or die "Can't open adj file: $afile\n";
open(ADV, ">$advfile") or die "Can't open adv file: $advfile\n";

for(my $sentence=0;$sentence<$#$inputArray;$sentence++){
    my $input= $$inputArray[$sentence];
     ## NP blocks
    while ( $input=~/\[\[ (.*?)\]\]/ism ) {
        my $tmp=$1;
        my $plain=stripAllTags($tmp);
        print NPS_BARE "$sentence:$plain\n";
        ## insert NP into an array of arrays
        my $phrase=convertToArray($tmp);
        my $l=@$phrase;
        ## print np by individual bits
        printNPBlock(\*NPS, \*ADJ, $phrase, 0, $l, $sentence);
        $input=~s/(\[\[.*?\]\])//ism ;
    }
    ## VP Blocks
    while ( $input=~/\(\((.*?)\)\)/ism ) {
        my $vpPhrase=convertToArray($1);
        my $l=@$vpPhrase;
        printVPBlock(\*VPS, \*ADV, $vpPhrase, $l, $sentence);
        $input=~s/(\(\(.*?\)\))//ism ;
    }
    while ( $input=~/\s+([a-z]+)_([A-Z]+)\s/ism ){
        my $word=$1;
        my $pos=$2;
        if($pos=~/JJ/){
            print ADJ "$sentence:$word\n";
        }
        elsif($pos=~/RB/){
            print ADV "$sentence:$word\n";
        }
        elsif($pos=~/VB/){
            print VPS "$sentence:$word\n";
        }
```

```
            $input=~s/\s+([a-z]+)_([A-Z]+)\s//ism ;
        }
    }
#### CLOSE FILES ####
    close NPS_BARE;
    close NPS;
    close VPS;
    close ADJ;
    close ADV;
}


#### SUBROUTINES #####
sub convertToArray {
    my $string = shift();
    my @myPhrase;
    my $counter=0;
    ## Remove starting and trailing white space
    $string =~ s/^\s+//;
    $string =~ s/\s+$//;
    ## get rid of odd punctuation
    $string =~ s/,_,//g;
    $string =~ s/-_://g;

    while($string =~ /^(.*?)_(\w+\$*)(\s*)/){
        $string =~ s/^(.*?)_(\w+\$*)(\s*)//;
        $myPhrase[$counter][0]=$1;
        $myPhrase[$counter][1]=$2;
        $myPhrase[$counter][2]=$3;
        ## Hack for E209
        if(($myPhrase[$counter][0]=~/^E\d+$/)
              || ($myPhrase[$counter][0]=~/^\+\s+E\d[\d,]+$/)){
            $myPhrase[$counter][0]="euros";
            $myPhrase[$counter][1]="NN";
            $myPhrase[$counter][2]=" ";
        }
        ## Hack for stray apostrophes
        if($myPhrase[$counter][1]!~/POS/
             && $myPhrase[$counter][0]=~/^\'(.*)$/){
            $myPhrase[$counter][0]=$1;
        }
```

```
        $counter++;
    }
    $myPhrase[$counter-1][2]="";

    return \@myPhrase;
}

sub printVPBlock {
    my ($fh, $advfh, $array, $length, $sentence) = @_;
    for(my $i=0;$i<$length;$i++){
        if(($array->[$i][1]=~/RB/ && $array->[$i][0]=~/^not$/)
                                || $array->[$i][0]=~/n\'t/){
            print $fh "***NEG***";
        }
        elsif($array->[$i][1]=~/RB/){
            print $advfh "$sentence:$array->[$i][0]\n";
        }
        elsif($array->[$i][1]=~/RP/ && $i>0){
            print $fh "$sentence:$array->[$i-1][0] $array->[$i][0]\n";
        }
        elsif($array->[$i][1]=~/VB/
                && $array->[$i][0]!~/^(have|\'ve|had|has|be|am|\'|are|\'re|
                        is|\'s|was|were|being|having|do|does|did|doing)$/
                && $array->[$i][0]!~/n\'t/ ){
            print $fh "$sentence:$array->[$i][0]\n";
        }
    }
}



sub printNPBlock {
    my ($fh, $adjfh, $array, $startPt, $endPt, $sentence) = @_;
    ## Local variables
    my $detsPresent=0;
    my $adjsPrinted=0;
    my $adjsPresent=0;
    my $flagNN=0;
    my $noOfWords=0;
    my $posBoundary=0;
    my $detBoundary=$startPt;
```

```
my $adjBoundary;
my $midNNboundary;

## print whole lot
for(my $i=$startPt;$i<$endPt;$i++){
    if($i==$startPt){
        print $fh "$sentence:";
    }
    print $fh "$array->[$i][0]$array->[$i][2]";
}
print $fh "\n";

## look for POS, pass this to printNPBlock
for(my $i=$startPt;$i<$endPt;$i++){
    if($array->[$i][1]=~/POS/){
        printNPBlock($fh, $adjfh, $array, $startPt, $i, $sentence);
        $posBoundary=$i+1;
    }
}
###################

## Print without determiners and mark determiners' end
for(my $i=$startPt;$i<$endPt;$i++){
    ## more than, greater than...
    if($i+1<@$array && $array->[$i][1]=~/(CD|PRP|RBR|RBS|JJS|JJR)/
                    && $array->[$i+1][0]=~/^(of|than)$/){
        $i++;
        ## only mark dets in the first group of dets
        if($flagNN==0){
            $detBoundary=$i+1;
        }
    }
    ## other dets and quantifiers
    elsif($array->[$i][1]=~/(CD|DT|PDT|PRP\$|WDT)/){
        if($flagNN==0){
            $detBoundary=$i+1;
        }
    }
    ## everything else
    else {
        $flagNN=1;
```

```
        ## print if have skipped some dets
        if($detBoundary>$startPt){
            if($i==$detBoundary){
                print $fh "$sentence:";
            }
            print $fh "$array->[$i][0]$array->[$i][2]";
            $detsPresent=1;
        }
    }
}
if($detsPresent==1){print $fh "\n";}
##################

## Print with adjectives only if had encounter DETS in prev section
## and mark AJDS end
$startPt=$detBoundary;
$adjBoundary=$detBoundary;
$flagNN=0;
for(my $i=$startPt;$i<$endPt;$i++){
    if($array->[$i][1]=~/(JJR|JJS|RBR|RBS|JJ|RB|CC|VB)/){
        ## Only print if hasn't been printed out in full before
        if($detsPresent==0){
            $adjsPrinted=1;
        }
        if($adjsPrinted>0){
            if($i==$startPt){
                print $fh "$sentence:";
            }
            print $fh "$array->[$i][0]$array->[$i][2]";
        }
        if($flagNN==0){
            $adjBoundary=$i+1;
            $adjsPresent=1;
        }
        if($1!~/CC/){
            print $adjfh "$sentence:$array->[$i][0]\n";
        }
    }
    else {
        $flagNN=1;
        if($adjsPrinted>0){
```

```
            if($i==$startPt){
                print $fh "$sentence:";
            }
            print $fh "$array->[$i][0]$array->[$i][2]";
        }
    }
}
if($adjsPrinted>0){print $fh "\n";}
##################

## Do the NNs from end of POS or end of ADJS
## whichever is furthest
if($posBoundary>$adjBoundary){
    $startPt=$posBoundary;
}
else {
    $startPt=$adjBoundary;
}
$midNNboundary=$startPt;
my $nounCount=0;
## do not print single nouns that have been printed
## CASE 1: lone word NPs
## CASE 2: DETs + lone noun
## Do print single nouns that haven't been printed
## CASE 1: POSSESSIVE + lone noun
## CASE 2: ADJs + lone noun
$noOfWords=$endPt-$startPt;
for(my $i=$startPt;$i<$endPt;$i++){
    if($array->[$i][1]=~/NN/){
        $nounCount++;
    }
}
if($nounCount==$noOfWords && $nounCount>1){
    printNounPermutations($fh, $array, $startPt, $endPt, $sentence);
}
elsif($noOfWords>1 || ($posBoundary>0 || $adjsPresent>0)){
    for(my $i=$startPt;$i<$endPt;$i++){
        if($i==$startPt){
            print $fh "$sentence:";
        }
        print $fh "$array->[$i][0]$array->[$i][2]";
```

```
            if($array->[$i][1]=~/NN/){
                $nounCount++;
            }
            ## If midNNboundary not changed already, change it.
            elsif($midNNboundary==$startPt){
                $midNNboundary=$i;
            }

            ## at the end of the loop
            if($i==$endPt-1){
                ## If the last element is not a noun
                ## print again ending before the last element
                if($array->[$i][1]!~/NN/){
                    $endPt=$i;
                    $i=$startPt-1;
                    $nounCount=0;
                }
                ## if have detected new midNNboundary (eg a JJ)
                ## start again from new midNNboundary point
                if($midNNboundary>$startPt && $midNNboundary<$endPt-1){
                    $startPt=$midNNboundary;
                    $i=$midNNboundary;
                    $nounCount=0;
                }

                ## if we are left with a string of nouns
                ## print the permutations
                elsif($nounCount==($endPt-$startPt)&&$nounCount>1){
                    print $fh "\n";
                    printNounPermutations($fh, $array,
                                          $startPt, $endPt, $sentence);
                }
                print $fh "\n";
            }
        }
    }
}

sub printNounPermutations {
    my ($fh, $array, $startPt, $endPt, $sentence) = @_;
    my $counter;
```

```
    for(my $i=$startPt;$i<$endPt;$i++){
        if($i==$startPt){
            print $fh "$sentence:";
        }
        print $fh "$array->[$i][0]\n";
        for(my $j=$i+1;$j<$endPt;$j++){
            if($j==$i+1){
                print $fh "$sentence:";
            }
            $counter=$i;
            while($counter<=$j){
                print $fh "$array->[$counter][0]";
                if($counter!=$j){
                    print $fh "$array->[$counter][2]";
                }
                $counter++;
            }
            print $fh "\n";
            if($i<$endPt-1){
                print $fh "$sentence:";
            }
        }
    }
}

sub stripAllTags {
    my $string = shift(@_);
    $string=~s/_.*?([\s|\'|-])/$1/g;
    $string=~s/_.*?$//g;
    $string =~ s/^\s+//;
    $string =~ s/\s+$//;
    return $string;
}

sub stripAllTagsAndBrackets {
    my $string = shift(@_);
    $string=~s/_.*?([\s|\'|-])/$1/g;
    $string=~s/_.*?$//g;
    $string =~ s/\<S\>//g;
    $string =~ s/(\(\(|\)\)|\[\[|\]\])//g;
    $string =~ s/(\s)\s+/$1/g;
```

```
    $string =~ s/^\s+//g;
    $string =~ s/\s+$//g;
    return $string;
}

sub readInputFile {
    my $filename = shift(@_);
    my $wholefile="";
    open(FILE, "$filename") or die "Can't open file: $filename\n";
    while(my $line=<FILE>){
        chomp $line;

        #### REMOVE OPENING BLURB ####
        if($line=~/\d\d\-\d\d\-\d\d\d\d \d\d\.\d\d/){
            next;
        }
         if($line=~/\d{6}TM\d+\.\d\d/){
             next;
         }
        if($line=~/^\d\.\s/){
            $line=join "", $line, ".";
        }

         if($line=~/DISC/){
             next;
         }
         if($line=~/RUNS=/){
             next;
         }
        if($line=~/\[(credit )?wires/){
            next;
        }
        $line=~s/\*+//;

        $wholefile=join " ", $wholefile, $line;
    }
    close FILE;
    print STDOUT $filename, " read\n";

    my @myInputArray=split /<\/S>/, $wholefile;
```

```
    return \@myInputArray;
}

sub printIndexedSentencesText {
    my $textArray = shift(@_);
    my $outfile = shift(@_);
    open(OUTFILE, ">$outfile") or die "Can't open file: $outfile\n";
    my $sentence;
    for(my $i=0;$i<$#$textArray;$i++){
        $sentence=$$textArray[$i];
        $sentence=&stripAllTagsAndBrackets($sentence);
        print OUTFILE "$i\t$sentence\n";
    }
    close OUTFILE;
}
```

C.3 Step 3 Relation Extraction source code

```
#!usr/bin/perl -w
use strict;
my ($WNdatafile, $in, $indir, $outdir, @allfiles, $ext, $seed, $seedfile);
my @pos=("n", "v", "a");

if($#ARGV!=3){
    die "Usage: perl $0 <WN data.all> <input dir> <output dir> <seed division: 0 or 1>\n";
}
else {
    $WNdatafile=$ARGV[0];
    $in=$ARGV[1];
    $outdir=$ARGV[2];
    $seed=$ARGV[3];
}

#### DATA STRUCTURES TO STORE VALUES ####
my %WN=();
#### RELATIONS I WANT ####
my @alwaysPrintRelations=("@", "%p", "%m", "%s", "#p", "#m", "#s", "*",
                          ">", "&", "^", "\$", "!", "=");
my @onlyPrintWhenInTextRelations=("#p", "#m", "#s", "-c", "-r", "-u", "+", "<", "\\");
my @followChainRelations=("@","*",">");

#### LOAD WN FILE ####
&readWNFiles(\%WN, $WNdatafile);
#     pointers => Hash of arrays:
#     keys are the pointers from the synset
#     values are arrays of the synsets pointed to

## GET WORDLIST FILES

if(-d $in){
    $indir=$in;
    opendir(DIR, "$indir") or die "couldn't open $indir: $!";
    @allfiles=grep {m <.*\.nps> }readdir DIR;
}
else {
    ($allfiles[0], $indir, $ext) = &fileparse($in);
    if(defined $ext){
```

```
        $allfiles[0]=$allfiles[0].$ext;
    }
 }

for my $j (0 .. $#allfiles){
    my @files=();
    $files[0]=join("/", $indir, $allfiles[$j]);
    ($files[1]= $files[0]) =~ s/.nps/.vps/;
    ($files[2]= $files[0]) =~ s/.nps/.adj/;
    my $outfile=join("/", $outdir, $allfiles[$j]);
    $outfile =~ s/.nps/.rel/;
    if($seed==1){
        ($seedfile=$outfile) =~ s/.rel/.seed/;
        $outfile =~ s/.rel/.ambrels/;
        open(SEED, "> $seedfile") or die "Can't open seedfile: $seedfile\n";
    }
    open(OUT, ">$outfile") or die "Can't open outfile: $outfile\n";
    my @phraseSynsets=();
    my @phraseTerms=();
    for(my $i=0;$i<=$#files;$i++){
        open(FILE, "$files[$i]") or die "Can't open $pos[$i] file: $files[$i]\n";
        while (defined (my $line = <FILE>)) {
            if($line!~/\d+:\w+/){
                next;
            }
            chomp $line;
            my @fields=split /:/, $line;
            my @synsets=();
            my $term=$fields[1];
            $term=~s/^\s+//;
            $term=~s/\s+$//;
            my ($senses, $anaWord, $anaSynset);
            if($term=~/^(i|me|my|you|your|yours|we|our|ours|they|their|theirs)$/i
                 || $term=~/^us$/){
                $anaWord=$term.":1";
                $anaSynset="00006026".$pos[$i];
                $senses="ana";
            }
            elsif($term=~/^(he|him|his)$/i){
                $anaWord=$term.":2";
                $anaSynset="09017703".$pos[$i];
```

```
    $senses="ana";
}
elsif($term=~/^(she|her|hers)$/i){
    $anaWord=$term.":2";
    $anaSynset="09013489".$pos[$i];
    $senses="ana";
}
elsif($term=~/^(it)$/i){
    $anaWord=$term.":1";
    $anaSynset="00001740".$pos[$i];
    $senses="ana";
}
else {
    $senses=`wn \"$term\" -syns$pos[$i] -o -s`;
}
if($senses ne ""){
    my $count=0;
    my $counter=1;
    my $sentence=$fields[0];
    $sentence=~s/\*{3}NEG\*{3}//;
    my $inText=1;
    if($senses ne "ana"){
        foreach my $sense (split /\n/, $senses){
            if($sense=~/^(Synonyms|Similarity)/){
                $counter=1;
            }
            if($sense=~/^\{(\d{8})\} (.*)/){
                my $word;
                $synsets[$count]=$1.$pos[$i];
                my $rest=$2;
                $rest=~s/\s*\(vs\. .+?\)//g;
                my @terms=split /, /, $rest;
                my $firstTwoLetters=substr($term, 0, 2);
                my $firstletter=substr($term, 0, 1);
                my ($tmp1, $tmp2, $index);
                for(my $x=0;$x<=$#terms;$x++){
                    if($terms[$x]=~/$term/i){
                        $index=$x;
                        last;
                    }
                    if($terms[$x]=~/($firstTwoLetters.+)\#$counter$/i){
```

```
                    $index=$x
                    }
                elsif($terms[$x]=~/($firstletter.+)\#$counter$/i){
                    $tmp1=$x
                    }
                elsif($terms[$x]=~/(.+)\#$counter$/i){
                    $tmp2=$x
                    }
            }
            if(!(defined($index))){
                if(defined($tmp1)){
                    $index=$tmp1;
                }
                else {
                    $index=$tmp2;
                }
            }
            $word=$term.":".$index;
            $phraseTerms[$sentence][$count]=$word;
            push (@{ $phraseSynsets[$sentence]{$word} }, $synsets[$count]);
            $count++;
            $counter++;
        }
    }
}
else{
    $synsets[$count]=$anaSynset;
    $phraseTerms[$sentence][$count]=$anaWord;
    push (@{ $phraseSynsets[$sentence]{$anaWord} }, $synsets[$count]);
}
## PRINT POLYSEMY RELS
for(my $a=0;$a<=$#synsets;$a++){
    for(my $b=$a+1;$b<=$#synsets;$b++){
      print OUT "pol\t$synsets[$a]:$synsets[$b]\t$sentence\t$inText
\t$inText\t$phraseTerms[$sentence][$a]\t$phraseTerms[$sentence][$b]\n";
      print OUT "pol\t$synsets[$b]:$synsets[$a]\t$sentence\t$inText
\t$inText\t$phraseTerms[$sentence][$b]\t$phraseTerms[$sentence][$a]\n";
    }
    if($seed==1 && $#synsets==0){
      &printRelations($synsets[$a], *SEED, $sentence, $inText,
                    $phraseTerms[$sentence][$a]);
```

```
                }
                else {
                  &printRelations($synsets[$a], *OUT, $sentence, $inText,
                                  $phraseTerms[$sentence][$a]);
                }
            }

        }
    }
    close(FILE);
}
close(OUT);
if($seed==1){
    close(SEED);
}
}

exit;

## SUB ROUTINES ##
sub printRelations {
    my $synset = shift();
    my $fh = shift();
    my $sentence = shift();
    my $inText = shift();
    my $term = shift();
    my $synNotInText=0;
    my $flag=0;
    ## print all relations wanted for all synsets
    foreach my $pointer (@alwaysPrintRelations){
        foreach my $pointee ( @{$WN{$synset}{pointers}{$pointer} } ){
            if(($pointer ne "&") ||  (defined $WN{$pointee}{pointers}{"!"}[0] ) ){
                if($pointer eq "@"){
                    print $fh "$pointer\t$pointee:$synset\t$sentence\t$synNotInText
                              \t$inText\tnull\t$term\n";
                }
                else{
                    print $fh "$pointer\t$synset:$pointee\t$sentence\t$inText
                              \t$synNotInText\t$term\tnull\n";
                }
            }
```

```
        }
    }
    if($synset=~/[nv]/ && $WN{$synset}{top}==1){
        print $fh "@\tTopx:$synset\t$sentence\t$synNotInText\t$inText\tnull\t$term\n";
    }

    ## print relations only required for synsets from text
    if($inText==1){
        foreach my $pointer (@onlyPrintWhenInTextRelations){
            foreach my $pointee ( @{$WN{$synset}{pointers}{$pointer} } ){
                print $fh "$pointer\t$synset:$pointee\t$sentence\t$inText
                        \t$synNotInText\t$term\tnull\n";
            }
        }
    }
    foreach my $pointer (@followChainRelations){
        foreach my $pointee ( @{$WN{$synset}{pointers}{$pointer} } ){
            if($pointer eq "*" && defined($WN{$pointee}{pointers}{"@"}[0])
                        && $WN{$pointee}{pointers}{"@"}[0] eq $synset ){
                next;
            }
            printRelations($pointee, $fh, $sentence, 0, "null");
        }
    }
}

## SUROUTINE TO LOAD WORDNET FILE ##
## JUST NEED POS, OFFSET, ALL POINTERS
sub readWNFiles {
    my $hashName = shift @_;
    my $fileName = shift @_;
    print STDERR "Reading file\n";
    open(FILE, "$fileName") or die "Couldn't open input file: $fileName\n";
    while (defined (my $line = <FILE>)) {
        chomp $line;
        ## For each synset
        if($line=~/^(\d{8})/){
            my $newRec={};
            my %pointerList=();
            my @fields=split(/\s/, $line);
            my $synset = $fields[0] . $fields[2];
```

```
            my $tmp=join("", "0x", $fields[3]);
            my $wdCount=oct $tmp;
            ## Get list of Words
            my $p_cntPosition=4+($wdCount*2);
            $newRec->{synno}=$synset;
            if($line!~/@/){
                $newRec->{top}=1;
            }
            else{
                $newRec->{top}=0;
            }
            if($line!~/~/){
                $newRec->{leafNode}=1;
            }
            else {
                $newRec->{leafNode}=0;
            }
            ## Get number of pointers
            my $p_count=$fields[$p_cntPosition];
            $newRec->{pcnt}=$p_count;
            ## Process pointers
            my $pointee;
            for(my $i=$p_cntPosition+1;$i<$p_cntPosition+1+($p_count*4);$i=$i+4){
                if($fields[$i+2]=~/[anv]/){
                    push @{ $pointerList{$fields[$i]} }, ($fields[$i+1] . $fields[$i+2]);
                }
            }
            $newRec->{pointers}= { %pointerList };
            $$hashName{ $newRec->{synno} } = $newRec;
        }
    }
    close FILE;
    print STDERR "File read\n";
}
```

C.4 WordNet Topographical File Derivation Source Code

```
#!usr/bin/perl -w
use strict;

my ($dictdir, @files, $index, %WN, %TopChildren);

if($#ARGV!=0){
    die "Usage: perl $0 <WN dict directory>\n";
}

else {
    $dictdir=$ARGV[0];
}

@files=();
$files[0]=$dictdir."data.noun";
$files[1]=$dictdir."data.verb";
$files[2]=$dictdir."data.adj";
$index=$dictdir."index.sense";
my %IndexHash=();
my %PolysemyWordHash=();
&readIndexSenseFile($index);
print STDERR "index file read\n";
for(my $f=0;$f<=$#files;$f++){
    open(FILE, "$files[$f]") or die "Couldn't open $f th input file: $files[$f]\n";
    #### DATA STRUCTURES TO STORE VALUES ####
    %WN=();
    %TopChildren=();
    my %Words=();

#########################################################
#### WN hash fields ####
#         Each record contains:
#         syyno => synset offset
#         pcnt => pointer count
#         words => list of words
#         wdCnt => number of words
#         leafnode => leaf=1 nonleaf=0
```

```
#         top => topnode=1 nontop=0
#         ## the keys are the pointers themselves ##
#         pointers => Hash of arrays:
#           keys are the pointers from the synset
#           values are arrays of the synsets pointed to
#           includes parents "@",
#           children "~", etc
#         depth => array of max and min depth
#         height => array of max and min height
#         descendants => number of descendants
#         ancestors => number of ancestors
#         sisters => number of sisters
#         BF2value => second degree branching facotr
#         inGroupWith => hash of hashes
#         E_LISTS => [ @E_List ],
#########################################################

    while (defined (my $line = <FILE>)) {
        chomp $line;
        ## For each synset
        if($line=~/^(\d{8})/){
            my $newRec={};
            my %pointerList=();
            my @fields=split(/\s/, $line);
            my $synset = $fields[0];
            my $tmp=join("", "0x", $fields[3]);
            my $wdCount=oct $tmp;
            ## Get list of Words
            my $p_cntPosition=4+($wdCount*2);
            my $list;
            for(my $i=4;$i<$p_cntPosition;$i=$i+2){
                my $wd=$fields[$i];
                ## for adjs really
                $wd=~s/\((a|p|ip)\)//g;
                if(defined($Words{$wd})){
                    $Words{$wd}++;
                }
                else {
                    $Words{$wd}=1;
                }
                my $key=$wd.$synset.$f;
```

```
        $key=lc $key;
        my $freq=$IndexHash{$key};
        my $lcWd=lc $wd;
        my $poly=$PolysemyWordHash{$lcWd};
        if($i==4){
            $list=$wd."%".$freq."%".$poly;
        }
        else{
            $list=$list.",".$wd."%".$freq."%".$poly;
        }
    }

    $newRec->{synno}=$synset;
    $newRec->{pos}=$fields[2];
    $newRec->{words}=$list;
    $newRec->{wdCnt}=$wdCount;

    if($line!~/@/){
        $newRec->{top}=1;
        $TopChildren{$synset}=();
    }
    else{
        $newRec->{top}=0;
    }
    if($line!~/ ~ /){
        $newRec->{leafNode}=1;
    }
    else {
        $newRec->{leafNode}=0;
    }
    ## Get number of pointers
    my $p_count=$fields[$p_cntPosition];
    my $p_counter=0;
    ## Process pointers
    for(my $i=$p_cntPosition+1;$i<$p_cntPosition+1+($p_count*4);$i=$i+4){
        push @{ $pointerList{$fields[$i]} }, $fields[$i+1];
        if($fields[$i] !~ /(~|#m|#p|#s|;c|;r|;u)/){
           $p_counter++;
              }
    }
    $p_count=~s/^0+//;
```

```
            $newRec->{pcnt}=$p_counter;
            $newRec->{pointers}= { %pointerList };
            $WN{ $newRec->{synno} } = $newRec;
        }
    }
    close FILE;
    print STDERR "File $files[$f] read\n";

    if($f==2){
        for my $syn (keys %WN){
            my $toPrint=getAdjOutputLine($syn);
            print STDOUT "$toPrint\n";
        }
    }
    else {
        foreach my $syn ( keys %TopChildren ) {
            getDepth($syn, 1, $syn);
            removeTopDuplicates($syn);
        }

        foreach my $syn ( keys %WN ){
            countSisters($syn);
        }

        for my $syn (keys %WN){
            my $toPrint=getOutputLine($syn);
            print STDOUT "$toPrint\n";
        }
    }
}
exit;

#### SUB ROUTINE TO PRINT VALS FOR CHILDREN OF EACH TOP TO SEPARATE FILE

sub getOutputLine {
    my $syn=shift @_;
    my $line = join("\t",
                    $WN{$syn}{pos},
                    $WN{$syn}{synno},
                    $WN{$syn}{pcnt},
                    $WN{$syn}{sisters},
```

```
                    $WN{$syn}{depth}[0],
                    $#{ $WN{$syn}{pointers}{"~"} }+1,
                    $WN{$syn}{leafNode},
                    $WN{$syn}{top},
                    $WN{$syn}{words},
                    );
    return $line;
}

sub getAdjOutputLine {
    my $syn=shift @_;
    my $line = join("\t",
                    "a",
                    $WN{$syn}{synno},
                    $WN{$syn}{pcnt},
                    $#{ $WN{$syn}{pointers}{"&"} }+1,
                    "0",
                    $#{ $WN{$syn}{pointers}{"!"} }+1,
                    "0",
                    "0",
                    $WN{$syn}{words},
                    );
    return $line;
}

#### COUNT SISTERS ####
#### LIKE BRANCHING FACTOR ALWAYS ADD 1 ####
#### HENCE, IF ONLY CHILD, SIS VALUE IS 1 ####
sub countSisters {
    my $syn=shift @_;
    if($WN{$syn}{top}==0){
        my $total=1;
        for my $i (0 .. $#{ $WN{$syn}{pointers}{"@"} } ){
            my $parent=$WN{$syn}{pointers}{"@"}[$i];
            $total+=$#{ $WN{$parent}{pointers}{"~"} }
        }
        $WN{$syn}{sisters}=$total;
    }
    else {
        $WN{$syn}{sisters}=scalar(keys %TopChildren);
    }
```

```
}

#### DEPTH
sub getDepth {
#    my ($syn, $down, $top, $index)=@_;
    my ($syn, $down, $top)=@_;
    push @{ $TopChildren{$top} }, $syn;
    if(!(defined ($WN{$syn}{depth}))){
        $WN{$syn}{depth}[0]=$down;
        $WN{$syn}{depth}[1]=$down;
    }
    elsif ($WN{$syn}{depth}[0]<$down){
        $WN{$syn}{depth}[0]=$down;
    }
    elsif ($WN{$syn}{depth}[1]>$down){
        $WN{$syn}{depth}[1]=$down;
    }
    if($WN{$syn}{leafNode}!=1){
        ### for each child in the array of children
        ### iterate down through the tree
        my $size = scalar ( @{ $WN{$syn}{pointers}{"~"} } );
        for(my $i=0;$i<$size;$i++ ) {
            getDepth($WN{$syn}{pointers}{"~"}[$i], $down+1, $top);
        }
    }
}

sub removeTopDuplicates{
    ### remove duplicates from Top children array
    my $top=shift @_;
    my $prev = 'nonesuch';
    my @in= sort @{$TopChildren{$top}};
    my @out = grep($_ ne $prev && ($prev = $_), @in);
    $TopChildren{$top}=[@out];
}

sub readIndexSenseFile{
    my $indexfile=shift();
    open(INDEX, "$indexfile") or die "Couldn't open input file: $indexfile\n";
    while (defined (my $line = <INDEX>)) {
        chomp $line;
```

```
        ## For each synset
        my $newRec={};
        my @fields=split(/\s/, $line);
        my $sensekey = $fields[0];
        my $synset = $fields[1];
        my $freq = $fields[3];
        my @keyfields=split(/%/, $line);
        my $word=$keyfields[0];
        my @lexSense=split(/:/,$keyfields[1]);
        my $pos=$lexSense[0];
        my $wdSynsetPos;
        if($pos==5){
            $pos=3;
        }
        if($pos<4){
            $pos=$pos-1;
            $wdSynsetPos=$word.$synset.$pos;
            $IndexHash{$wdSynsetPos}=$freq;
        }
        if(!(defined($PolysemyWordHash{$word}))){
            $PolysemyWordHash{$word}=0;
        }
        else {
            $PolysemyWordHash{$word}++;
        }
    }
    close INDEX;
}
```

C.5 Input and Output File Sample

C.5.1 Step 1 input

1. Strike by NI firefighters under way

2. Firefighters' strike is under way

The British Army has taken over firefighting duties in the North at the start of an eight-day strike by fire crews. Almost 2,000 full-time firefighters have left their posts and pickets are being put on fire stations. The pay strike by firefighters throughout Britain and Northern Ireland began at 9am this morning after overnight talks failed to avert it. The Fire Brigades' Union has accused the British government of vetoing an agreement that could have prevented the action. The union said it had worked out a deal with employers that would have given firefighters a 16 pay increase, but the government had refused to help fund it.

C.5.2 Step 1 output = Step 2 input

```
<S>1_LS._.</S> <S>[[ Strike_NN ]] by_IN [[ NI_NNP firefighters_NNS ]]
under_IN [[ way_NN
2_CD ]]._.</S> <S>[[ Firefighters_NNS'_POS strike_NN ]] (( is_VBZ ))
under_IN [[ way_NN
The_NNP British_NNP Army_NNP ]] (( has_VBZ taken_VBN )) over_IN
[[ firefighting_JJ duties_NNS ]] in_IN [[ the_DT North_NNP ]] at_IN
[[ the_DT start_NN ]] of_IN [[ an_DT eight-day_JJ strike_NN ]] by_IN
[[ fire_NN crews_NNS ]]._.</S>
<S>Almost_RB [[ 2,000_CD full-time_JJ firefighters_NNS ]]
(( have_VBP left_VBN )) [[ their_PRP\$ posts_NNS ]] and_CC
[[ pickets_NNS ]] (( are_VBP being_VBG put_VBN )) on_IN
[[ fire_NN stations_NNS ]]._.</S>
<S>[[ The_DT pay_NN strike_NN ]] by_IN [[ firefighters_NNS ]]
throughout_IN [[ Britain_NNP ]] and_CC [[ Northern_NNP Ireland_NNP ]]
(( began_VBD )) at_IN [[ 9am_CD ]] [[ this_DT morning_NN ]] after_IN
[[ overnight_JJ talks_NNS ]] (( failed_VBD )) (( to_TO avert_VB ))
[[ it_PRP ]]._.</S>
<S>[[ The_DT Fire_NNP Brigades_NNPS'_POS Union_NNP ]]
(( has_VBZ accused_VBN )) [[ the_DT British_JJP government_NN ]]
of_IN vetoing_VBG [[ an_DT agreement_NN ]] [[ that_WDT ]]
(( could_MD have_VB prevented_VBN )) [[ the_DT action_NN ]]._.</S>
<S>[[ The_DT union_NN ]] (( said_VBD )) [[ it_PRP ]]
(( had_VBD worked_VBN )) out_IN [[ a_DT deal_NN ]] with_IN
[[ employers_NNS ]] [[ that_WDT ]] (( would_MD have_VB given_VBN ))
[[ firefighters_NNS ]] [[ a_DT 16_CD ]]%_SYM [[ pay_NN increase_NN ]]
,_, but_CC [[ the_DT government_NN ]] (( had_VBD refused_VBN ))
(( to_TO help_VB )) [[ fund_NN ]] [[ it_PRP ]]._.</S>
```

C.5.3 Step 2 output = Step 3 input

The output of Step 2 is as follows:

SentenceNumber:TermExtracted

- Nouns (221102_AMDfirefighters8.nps)

```
1:NI
1:NI firefighters
1:Strike
1:firefighters
1:way
1:way 2
2:Army
2:British
2:British Army
2:Firefighters
2:Firefighters' strike
2:North
2:The
2:The British
2:The British Army
2:an eight-day strike
2:crews
2:duties
2:eight-day strike
2:fire
2:fire crews
2:firefighting duties
2:start
2:strike
2:the North
2:the start
2:way
```

2:way The

2:way The British

2:way The British Army

3:2,000 full-time firefighters

3:fire

3:fire stations

3:firefighters

3:full-time firefighters

3:pickets

3:posts

3:stations

3:their posts

4:9am

4:Britain

4:Ireland

4:Northern

4:Northern Ireland

4:The pay strike

4:firefighters

4:it

4:morning

4:overnight talks

4:pay

4:pay strike

4:strike

4:talks

4:this morning

5:Brigades

5:British government

5:Fire

5:Fire Brigades

5:Fire Brigades' Union

5:The Fire Brigades

5:The Fire Brigades' Union

5:Union

5:action

5:agreement

5:an agreement

5:government

5:that

5:the British government

5:the action

6:The union

6:a 16

6:a deal

6:deal

6:employers

6:firefighters

6:fund

6:government

6:increase

6:it

6:pay

6:pay increase

6:that

6:the government

6:union

- Verbs (221102_AMDfirefighters8.vps)

2:taken

3:left

3:put

4:avert

4:began

4:failed

5:accused

5:prevented

6:given

6:help

6:refused

6:said

6:worked

- Adjectives (221102_AMDfirefighters8.adj)

2:eight-day

2:firefighting

3:full-time

4:overnight

5:British

C.5.4 Step 3 output = Step 4 input

This is a portion of the relations that are output at Step 3.

```
@ 13788338n:13808178n 1 0 1 null NI:0
#s 13808178n:13836665n 1 1 0 NI:0 null
#s 13808178n:13842018n 1 1 0 NI:0 null
#s 13808178n:13849636n 1 1 0 NI:0 null
#s 13808178n:14189789n 1 1 0 NI:0 null
#s 13808178n:13836665n 1 1 0 NI:0 null
#s 13808178n:13842018n 1 1 0 NI:0 null
#s 13808178n:13849636n 1 1 0 NI:0 null
#s 13808178n:14189789n 1 1 0 NI:0 null
+ 13808178n:01355211v 1 1 0 NI:0 null
@ 13785984n:13788338n 1 0 0 null null
@ 00017572n:13785984n 1 0 0 null null
%p 13785984n:13782418n 1 0 0 null null
@ 00001740n:00017572n 1 0 0 null null
@ Topx:00001740n 1 0 0 null null
pol 01173759n:00919554n 1 1 1 Strike:0 Strike:0
pol 00919554n:01173759n 1 1 1 Strike:0 Strike:0
pol 01173759n:00101927n 1 1 1 Strike:0 Strike:0
pol 00101927n:01173759n 1 1 1 Strike:0 Strike:0
pol 01173759n:06956656n 1 1 1 Strike:0 Strike:1
pol 06956656n:01173759n 1 1 1 Strike:1 Strike:0
pol 01173759n:00177371n 1 1 1 Strike:0 Strike:0
pol 00177371n:01173759n 1 1 1 Strike:0 Strike:0
pol 01173759n:00059106n 1 1 1 Strike:0 Strike:3
pol 00059106n:01173759n 1 1 1 Strike:3 Strike:0
@ 01172321n:01173759n 1 0 1 null Strike:0
+ 01173759n:02339744v 1 1 0 Strike:0 null
```

C.5.5 Step 4 input: WordNet topographical features file

This is a short excerpt from the file derived from WordNet synsets, synonymous word lists and topographical features. See appendix C.4 for the source code of the perl script to derive this file from the WordNet database files.

Each record, i.e. line, relates to a unique synset in WordNet. The record consists of 9 tab-delimited fields which are as follows:

1. Part of speech
2. Synset offset: a unique identifier for each synset within each part of speech category of WordNet
3. Total number of connections from that synset to others
4. Number of coordinate terms for that Synset (number of sisters)
5. Maximum depth of the Synset in WordNet
6. Synset branching factor (number of children)
7. LeafNode: binary field indicating if synset is a leaf node or not
8. TopNode: binary field indicating if synset is a top node in WordNet or not
9. Synonymous word list

   This field itself consists of three fields for each of the synonymous words, delimited by the percentage (%) symbol:

   (a) Word form

   (b) Word frequency, according to WordNet sense tagged corpus

   (c) Word polysemy count: the number of synsets to which this word belongs

```
n 07054143 2 2 5 0 1 0 electricity%1%2
n 11707751 1 25 7 0 1 0 Hemerocallidaceae%0%0,family_Hemerocallidaceae%0%0
n 03266377 1 2 12 0 1 0 fowling_piece%0%0
n 02049333 2 359 6 0 1 0 Selenarctos%0%0,genus_Selenarctos%0%0
n 11322732 1 87 11 0 1 0 lizard_orchid%0%0,Himantoglossum_hircinum%0%0
n 14422506 1 1 7 0 1 0 luminous_flux%0%0
n 12303370 2 135 6 0 1 0 Fomes%0%0,genus_Fomes%0%0
```

```
n 10721539 1 16 5 0 1 0 greenhouse_effect%0%0
n 00654095 1 3 8 1 0 0 enema%0%0,clyster%0%0
n 11045372 2 233 6 0 1 0 genus_Pistia%0%0
n 09189598 6 40 7 1 0 0 asserter%0%0,declarer%0%1,affirmer%0%0,asseverator%0%0,avower%0%1
n 12681082 1 4 8 0 1 0 bowel_movement%0%0,movement%0%10,bm%0%1
n 04492427 3 6 7 0 1 0 bewitchery%0%0,beguilement%0%1,animal_magnetism%0%0
n 03318319 1 2 7 0 1 0 Golconda%0%0
n 09041946 1 5 11 0 1 0 Arikara%0%1,Aricara%0%1
n 07473247 3 92 3 0 1 0 battery%2%6
n 07889876 1 92 3 0 1 0 batch%1%3
n 08986326 1 77 8 0 1 0 Babar%0%0
n 07857947 1 3 6 0 1 0 mercantilism%0%1,mercantile_system%0%0
v 00777368 8 10 5 7 0 0 authorize%32%1,authorise%0%1,pass%0%42,clear%2%45
v 00279987 1 16 3 1 0 0 redden%0%2
v 02286040 1 2 3 0 1 0 distrain%0%2
v 02017278 1 19 2 0 1 0 slip_away%1%1,steal_away%1%0,sneak_away%0%0,sneak_off%2%0
v 02574709 0 554 1 0 1 1 go%14%34
v 01687501 3 2 4 0 1 0 photostat%0%2
v 02303587 1 2 10 0 1 0 bull%1%14,bull_through%0%0
a 02262863 1 1 0 0 0 0 premium%1%3
a 00139578 1 1 0 0 0 0 accessible%0%3
a 00187598 1 1 0 0 0 0 acquirable%0%0
a 01751315 1 1 0 0 0 0 spiked%1%2
a 02006171 1 1 0 0 0 0 certifiable%0%1,certified%0%3
a 02336718 15 11 0 1 0 0 untidy%0%1
a 02899542 1 0 0 0 0 0 Cyrillic%0%1
a 00335396 1 1 0 0 0 0 diffident%1%1,shy%3%6,timid%0%4,unsure%0%2
```

Appendix D

Gold Standard Experiment Materials

D.1 Talk aloud protocol materials

(D.1) FLOODING

Cork is one of the worst areas affected by the overnight rain. There is flooding on the Ballyhooley Road and in the Ballyvolane and Togher areas of the city. There is also flooding on the N20 immediately outside Mallow, on the Limerick side at Kilknockin as well as on the back road between the Glanmire Road and Glashaboy Waterworks. Many houses in these areas are also flooded, as are some in Blackpool in Cork City. Gardai in the city have appealed to motorists to slow down. They also say the City Centre is still clear of flooding but high tide is at 10.35 and conditions could worsen. The main Limerick to Cork road also has flooding. And in the Parteen area of Limerick City, back roads have been flooded. In Kerry, no roads are impassible but water on roads is making driving conditions dangerous in places.

****

The Flood Tribunal has heard that an Independent Dublin councillor, John O'Halloran, has admitted receiving 2,500 from the former lobbyist, Frank Dunlop, during the 1996 Dublin West By-election. In a statement being read at the enquiry, Mr O'Halloran said he personally solicited this money as a political donation and that it was not connected with any rezoning motion or planning matter. The Tribunal heard Brian O'Halloran - one of the former owners of the Carrickmines site - say he was never approached for money by any politician for money in relation to having the land rezoned, nor did he offer any inducement. Two other co-owners, Gerard Kilcoyne and Profesoor Austin Darragh, also said that they said they were never approached for money by any politician in relation to having the land rezoned nor did they offer any inducement.

- **Identical**
- **Very similar**
- **Quite similar**
- **Quite different**
- **Very different**
- **Completely different**

(D.2) FLOODING

Cork is one of the worst areas affected by the overnight rain. There is flooding on the Ballyhooley Road and in the Ballyvolane and Togher areas of the city. There is also flooding on the N20 immediately outside Mallow, on the Limerick side at Kilknockin as well as on the back road between the Glanmire Road and Glashaboy Waterworks. Many houses in these areas are also flooded, as are some in Blackpool in Cork City. Gardai in the city have appealed to motorists to slow down. They also say the City Centre is still clear of flooding but high tide is at 10.35 and conditions could worsen. The main Limerick to Cork road also has flooding. And in the Parteen area of Limerick City, back roads have been flooded. In Kerry, no roads are impassible but water on roads is making driving conditions dangerous in places.

****

The Flood Tribunal has heard that an Independent Dublin councillor, John O'Halloran, has admitted receiving 2,500 from the former lobbyist, Frank Dunlop, during the 1996 Dublin West By-election. In a statement being read at the enquiry, Mr O'Halloran said he personally solicited this money as a political donation and that it was not connected with any rezoning motion or planning matter. The Tribunal heard Brian O'Halloran - one of the former owners of the Carrickmines site - say he was never approached for money by any politician for money in relation to having the land rezoned, nor did he offer any inducement. Two other co-owners, Gerard Kilcoyne and Profesoor Austin Darragh, also said that they said they were never approached for money by any politician in relation to having the land rezoned nor did they offer any inducement.

- Are these two stories the same?
- Are these two stories about the same event?
- Are these two stories about the same type of event?
- Are these two stories on the same topic?
- Are these two stories the same genre?
- Are these two stories similar?

(D.3) FLOODING

Cork is one of the worst areas affected by the overnight rain. There is flooding on the Ballyhooley Road and in the Ballyvolane and Togher areas of the city. There is also flooding on the N20 immediately outside Mallow, on the Limerick side at Kilknockin as well as on the back road between the Glanmire Road and Glashaboy Waterworks. Many houses in these areas are also flooded, as are some in Blackpool in Cork City. Gardai in the city have appealed to motorists to slow down. They also say the City Centre is still clear of flooding but high tide is at 10.35 and conditions could worsen. The main Limerick to Cork road also has flooding. And in the Parteen area of Limerick City, back roads have been flooded. In Kerry, no roads are impassible but water on roads is making driving conditions dangerous in places.

****

The Flood Tribunal has heard that an Independent Dublin councillor, John O'Halloran, has admitted receiving 2,500 from the former lobbyist, Frank Dunlop, during the 1996 Dublin West By-election. In a statement being read at the enquiry, Mr O'Halloran said he personally solicited this money as a political donation and that it was not connected with any rezoning motion or planning matter. The Tribunal heard Brian O'Halloran - one of the former owners of the Carrickmines site - say he was never approached for money by any politician for money in relation to having the land rezoned, nor did he offer any inducement. Two other co-owners, Gerard Kilcoyne and Profesoor Austin Darragh, also said that they said they were never approached for money by any politician in relation to having the land rezoned nor did they offer any inducement.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Identical	..	..	..	..	..	..	..	..	**Totally different**

(D.4) 1. Flood hears councillors' denials 2. Cash for votes denied at Tribunal The Flood Tribunal has heard formal written denials from two Dublin County Councillors that they accepted money in return for voting in favour of rezoning motions. Councillor Tony Fox denied ever receiving any money from Frank Dunlop. Councillor Colm McGrath said he was given a number of unconditional political donations from the lobbyist, but these were not connected with the planning process. Both men were named yesterday in an opening statement to the Tribunal as accepting bribes in return for their votes. The Tribunal is now hearing the correspondence between it and former Fine Gael senator Liam Cosgrave.
****

The Flood Tribunal has heard that an Independent Dublin councillor, John O'Halloran, has admitted receiving 2,500 from the former lobbyist, Frank Dunlop, during the 1996 Dublin West By-election. In a statement being read at the enquiry, Mr O'Halloran said he personally solicited this money as a political donation and that it was not connected with any rezoning motion or planning matter. The Tribunal heard Brian O'Halloran - one of the former owners of the Carrickmines site - say he was never approached for money by any politician for money in relation to having the land rezoned, nor did he offer any inducement. Two other co-owners, Gerard Kilcoyne and Profesoor Austin Darragh, also said that they said they were never approached for money by any politician in relation to having the land rezoned nor did they offer any inducement.

- **Identical**
- **Very similar**
- **Quite similar**
- **Quite different**
- **Very different**
- **Completely different**

(D.5) 1. Flood hears councillors' denials 2. Cash for votes denied at Tribunal The Flood Tribunal has heard formal written denials from two Dublin County Councillors that they accepted money in return for voting in favour of rezoning motions. Councillor Tony Fox denied ever receiving any money from Frank Dunlop. Councillor Colm McGrath said he was given a number of unconditional political donations from the lobbyist, but these were not connected with the planning process. Both men were named yesterday in an opening statement to the Tribunal as accepting bribes in return for their votes. The Tribunal is now hearing the correspondence between it and former Fine Gael senator Liam Cosgrave.
****

The Flood Tribunal has heard that an Independent Dublin councillor, John O'Halloran, has admitted receiving 2,500 from the former lobbyist, Frank Dunlop, during the 1996 Dublin West By-election. In a statement being read at the enquiry, Mr O'Halloran said he personally solicited this money as a political donation and that it was not connected with any rezoning motion or planning matter. The Tribunal heard Brian O'Halloran - one of the former owners of the Carrickmines site - say he was never approached for money by any politician for money in relation to having the land rezoned, nor did he offer any inducement. Two other co-owners, Gerard Kilcoyne and Profesoor Austin Darragh, also said that they said they were never approached for money by any politician in relation to having the land rezoned nor did they offer any inducement.

- Are these two stories the same?
- Are these two stories about the same event?
- Are these two stories about the same type of event?
- Are these two stories on the same topic?
- Are these two stories the same genre?
- Are these two stories similar?

(D.6) 1. Flood hears councillors' denials 2. Cash for votes denied at Tribunal The Flood Tribunal has heard formal written denials from two Dublin County Councillors that they accepted money in return for voting in favour of rezoning motions. Councillor Tony Fox denied ever receiving any money from Frank Dunlop. Councillor Colm McGrath said he was given a number of unconditional political donations from the lobbyist, but these were not connected with the planning process. Both men were named yesterday in an opening statement to the Tribunal as accepting bribes in return for their votes. The Tribunal is now hearing the correspondence between it and former Fine Gael senator Liam Cosgrave.

****

The Flood Tribunal has heard that an Independent Dublin councillor, John O'Halloran, has admitted receiving 2,500 from the former lobbyist, Frank Dunlop, during the 1996 Dublin West By-election. In a statement being read at the enquiry, Mr O'Halloran said he personally solicited this money as a political donation and that it was not connected with any rezoning motion or planning matter. The Tribunal heard Brian O'Halloran - one of the former owners of the Carrickmines site - say he was never approached for money by any politician for money in relation to having the land rezoned, nor did he offer any inducement. Two other co-owners, Gerard Kilcoyne and Profesoor Austin Darragh, also said that they said they were never approached for money by any politician in relation to having the land rezoned nor did they offer any inducement.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Identical	..	..	..	..	..	..	..	..	**Totally different**

(D.7) budget - social welfare Finance Minister Charlie McCreevy told the Dail that the total cost of Social Welfare improvements in today's Budget is E501 million in 2003 and E530 million in a full year. He said the maximum weekly personal rate for all old age pensions will be increased by E10 from the first week of January 2003. Child Benefit will be increased by E8 a month for each of the first and second children to E125.60 a month and by E10 for each of the third and subsequent children to E157.30 a month. The increases will be effective from next April and will cost E78.62 million in 2003 and E104.83 in a full year. From June 2003, the Respite Care Grant will be increased by E100 to E735.
****

1800McCULLAGH The Minister for Finance, Charlie McCreevy, has announced public spending increases in this afternoon's budget, but he firmly warned at the outset of his speech that the economic situation had worsened because of the weakness of the global economy. There will be public spending increases, but there are also increases in indirect taxation. While income tax remains unaffected, stamp duties are to go up significantly from midnight, with a large increase on the rate charged on credit cards. Duties on ATM cards go up to ten euro a year, and the duty on combined ATM/laser cards will be 20 euro a year. Petrol and beer have been spared, but the price of cigarettes, diesel, spirits and alcopops goes up. Fine Gael's Finance spokesman, Richard Bruton, described the budget as "genetically flawed".

- **Identical**
- **Very similar**
- **Quite similar**
- **Quite different**
- **Very different**
- **Completely different**

(D.8) budget - social welfare Finance Minister Charlie McCreevy told the Dail that the total cost of Social Welfare improvements in today's Budget is E501 million in 2003 and E530 million in a full year. He said the maximum weekly personal rate for all old age pensions will be increased by E10 from the first week of January 2003. Child Benefit will be increased by E8 a month for each of the first and second children to E125.60 a month and by E10 for each of the third and subsequent children to E157.30 a month. The increases will be effective from next April and will cost E78.62 million in 2003 and E104.83 in a full year. From June 2003, the Respite Care Grant will be increased by E100 to E735.
****

1800McCULLAGH The Minister for Finance, Charlie McCreevy, has announced public spending increases in this afternoon's budget, but he firmly warned at the outset of his speech that the economic situation had worsened because of the weakness of the global economy. There will be public spending increases, but there are also increases in indirect taxation. While income tax remains unaffected, stamp duties are to go up significantly from midnight, with a large increase on the rate charged on credit cards. Duties on ATM cards go up to ten euro a year, and the duty on combined ATM/laser cards will be 20 euro a year. Petrol and beer have been spared, but the price of cigarettes, diesel, spirits and alcopops goes up. Fine Gael's Finance spokesman, Richard Bruton, described the budget as "genetically flawed".

- Are these two stories the same?
- Are these two stories about the same event?
- Are these two stories about the same type of event?
- Are these two stories on the same topic?
- Are these two stories the same genre?
- Are these two stories similar?

(D.9) budget - social welfare Finance Minister Charlie McCreevy told the Dail that the total cost of Social Welfare improvements in today's Budget is E501 million in 2003 and E530 million in a full year. He said the maximum weekly personal rate for all old age pensions will be increased by E10 from the first week of January 2003. Child Benefit will be increased by E8 a month for each of the first and second children to E125.60 a month and by E10 for each of the third and subsequent children to E157.30 a month. The increases will be effective from next April and will cost E78.62 million in 2003 and E104.83 in a full year. From June 2003, the Respite Care Grant will be increased by E100 to E735.
****

1800McCULLAGH The Minister for Finance, Charlie McCreevy, has announced public spending increases in this afternoon's budget, but he firmly warned at the outset of his speech that the economic situation had worsened because of the weakness of the global economy. There will be public spending increases, but there are also increases in indirect taxation. While income tax remains unaffected, stamp duties are to go up significantly from midnight, with a large increase on the rate charged on credit cards. Duties on ATM cards go up to ten euro a year, and the duty on combined ATM/laser cards will be 20 euro a year. Petrol and beer have been spared, but the price of cigarettes, diesel, spirits and alcopops goes up. Fine Gael's Finance spokesman, Richard Bruton, described the budget as "genetically flawed".

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Identical	..	..	..	..	..	..	..	..	**Totally different**

(D.10) 1. Strike by NI firefighters begins today 2. Firefighters' strike begins today The British Army has taken over firefighting duties in the North at the start of an eight-day strike by fire crews. Almost 2,000 full-time firefighters have left their posts and pickets are being put on fire stations. The pay strike by firefighters throughout Britain and Northern Ireland began at 9am this morning after overnight talks failed to avert it. The Fire Brigades' Union has accused the British government of vetoing an agreement that could have prevented the action. The union said it had worked out a deal with employers that would have given firefighters a 16% pay increase, but the government had refused to help fund it.
****

1. Strike by NI firefighters under way 2. Firefighters' strike is under way The British Army has taken over firefighting duties in the North at the start of an eight-day strike by fire crews. Almost 2,000 full-time firefighters have left their posts and pickets are being put on fire stations. The pay strike by firefighters throughout Britain and Northern Ireland began at 9am this morning after overnight talks failed to avert it. The Fire Brigades' Union has accused the British government of vetoing an agreement that could have prevented the action. The union said it had worked out a deal with employers that would have given firefighters a 16% pay increase, but the government had refused to help fund it.

- **Identical**
- **Very similar**
- **Quite similar**
- **Quite different**
- **Very different**
- **Completely different**

(D.11) 1. Strike by NI firefighters begins today 2. Firefighters' strike begins today The British Army has taken over firefighting duties in the North at the start of an eight-day strike by fire crews. Almost 2,000 full-time firefighters have left their posts and pickets are being put on fire stations. The pay strike by firefighters throughout Britain and Northern Ireland began at 9am this morning after overnight talks failed to avert it. The Fire Brigades' Union has accused the British government of vetoing an agreement that could have prevented the action. The union said it had worked out a deal with employers that would have given firefighters a 16% pay increase, but the government had refused to help fund it.

****

1. Strike by NI firefighters under way 2. Firefighters' strike is under way The British Army has taken over firefighting duties in the North at the start of an eight-day strike by fire crews. Almost 2,000 full-time firefighters have left their posts and pickets are being put on fire stations. The pay strike by firefighters throughout Britain and Northern Ireland began at 9am this morning after overnight talks failed to avert it. The Fire Brigades' Union has accused the British government of vetoing an agreement that could have prevented the action. The union said it had worked out a deal with employers that would have given firefighters a 16% pay increase, but the government had refused to help fund it.

- Are these two stories the same?
- Are these two stories about the same event?
- Are these two stories about the same type of event?
- Are these two stories on the same topic?
- Are these two stories the same genre?
- Are these two stories similar?

(D.12) 1. Strike by NI firefighters begins today 2. Firefighters' strike begins today The British Army has taken over firefighting duties in the North at the start of an eight-day strike by fire crews. Almost 2,000 full-time firefighters have left their posts and pickets are being put on fire stations. The pay strike by firefighters throughout Britain and Northern Ireland began at 9am this morning after overnight talks failed to avert it. The Fire Brigades' Union has accused the British government of vetoing an agreement that could have prevented the action. The union said it had worked out a deal with employers that would have given firefighters a 16% pay increase, but the government had refused to help fund it.

****

1. Strike by NI firefighters under way 2. Firefighters' strike is under way The British Army has taken over firefighting duties in the North at the start of an eight-day strike by fire crews. Almost 2,000 full-time firefighters have left their posts and pickets are being put on fire stations. The pay strike by firefighters throughout Britain and Northern Ireland began at 9am this morning after overnight talks failed to avert it. The Fire Brigades' Union has accused the British government of vetoing an agreement that could have prevented the action. The union said it had worked out a deal with employers that would have given firefighters a 16% pay increase, but the government had refused to help fund it.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Identical	..	..	..	..	..	..	..	..	**Totally different**

(D.13) 1. Sharon unveils proposals for Palestinian state 2. Proposals for a Palestinian state The Israeli Prime Minister, Ariel Sharon, has unveiled, for the first time, his detailed proposals for the creation of a Palestinian state. He said the proposals were within the framework announced earlier this year by US President, George W Bush. However, among the conditions set by Mr Sharon was a complete replacement of the current Palestinian leadership, headed by Yasser Arafat. The proposal was immediately rejected by the Palestinian chief negotiator, Saeb Erekat. Mr Sharon proposes the setting up of a Palestinian state within parts of the West Bank and Gaza, with provisional boundaries by next year and definitive borders by 2005. ********** 1. Irishwoman wounded in Jenin gunfight 2. Irishwoman hurt in Jenin gunfight An Irishwoman has been wounded in a gun battle between Israeli forces and Palestinian gunmen in a refugee camp in the northern West Bank city of Jenin. Caoimhe Butterly is a United Nations observer. In a separate gun battle in the same camp earlier, a British UN worker, Ian Hook, was shot dead. Earlier it was reported that a Palestinian policeman and an Israeli soldier were killed in separate incidents near Gaza City. Israeli forces launched incursions into the Gaza Strip and the West Bank city of Bethlehem overnight. In Bethlehem, troops sealed off the Church of the Nativity to prevent militants taking refuge in the compound as they did during Israel's reoccupation of the city in April.

- **Identical**
- **Very similar**
- **Quite similar**
- **Quite different**
- **Very different**
- **Completely different**

(D.14) 1. Sharon unveils proposals for Palestinian state 2. Proposals for a Palestinian state The Israeli Prime Minister, Ariel Sharon, has unveiled, for the first time, his detailed proposals for the creation of a Palestinian state. He said the proposals were within the framework announced earlier this year by US President, George W Bush. However, among the conditions set by Mr Sharon was a complete replacement of the current Palestinian leadership, headed by Yasser Arafat. The proposal was immediately rejected by the Palestinian chief negotiator, Saeb Erekat. Mr Sharon proposes the setting up of a Palestinian state within parts of the West Bank and Gaza, with provisional boundaries by next year and definitive borders by 2005. ********** 1. Irishwoman wounded in Jenin gunfight 2. Irishwoman hurt in Jenin gunfight An Irishwoman has been wounded in a gun battle between Israeli forces and Palestinian gunmen in a refugee camp in the northern West Bank city of Jenin. Caoimhe Butterly is a United Nations observer. In a separate gun battle in the same camp earlier, a British UN worker, Ian Hook, was shot dead. Earlier it was reported that a Palestinian policeman and an Israeli soldier were killed in separate incidents near Gaza City. Israeli forces launched incursions into the Gaza Strip and the West Bank city of Bethlehem overnight. In Bethlehem, troops sealed off the Church of the Nativity to prevent militants taking refuge in the compound as they did during Israel's reoccupation of the city in April.

- Are these two stories the same?
- Are these two stories about the same event?
- Are these two stories about the same type of event?
- Are these two stories on the same topic?
- Are these two stories the same genre?
- Are these two stories similar?

(D.15) 1. Sharon unveils proposals for Palestinian state 2. Proposals for a Palestinian state The Israeli Prime Minister, Ariel Sharon, has unveiled, for the first time, his detailed proposals for the creation of a Palestinian state. He said the proposals were within the framework announced earlier this year by US President, George W Bush. However, among the conditions set by Mr Sharon was a complete replacement of the current Palestinian leadership, headed by Yasser Arafat. The proposal was immediately rejected by the Palestinian chief negotiator, Saeb Erekat. Mr Sharon proposes the setting up of a Palestinian state within parts of the West Bank and Gaza, with provisional boundaries by next year and definitive borders by 2005.

****

1. Irishwoman wounded in Jenin gunfight 2. Irishwoman hurt in Jenin gunfight An Irishwoman has been wounded in a gun battle between Israeli forces and Palestinian gunmen in a refugee camp in the northern West Bank city of Jenin. Caoimhe Butterly is a United Nations observer. In a separate gun battle in the same camp earlier, a British UN worker, Ian Hook, was shot dead. Earlier it was reported that a Palestinian policeman and an Israeli soldier were killed in separate incidents near Gaza City. Israeli forces launched incursions into the Gaza Strip and the West Bank city of Bethlehem overnight. In Bethlehem, troops sealed off the Church of the Nativity to prevent militants taking refuge in the compound as they did during Israel's reoccupation of the city in April.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Identical	..	..	..	..	..	..	..	..	**Totally different**

(D.16) 1. New tourism body established 2. New tourism authority established The first head of a new tourism body established under the terms of the Good Friday agreement has been named. Shaun Quinn will be chief executive of the National Tourism Development Authority that will form part of the Tourism Ireland agency. He will take up the position in time for the start of the 2003 tourism season. The authority is to take over responsibility for developing the industry from Bord Failte, which will now concentrate on specialist functions, including sporting events such as the 2006 Ryder Cup golf. The NTDA will operate under the name Failte Ireland. Legislation to set it up was introduced last month in the Dail. Mr Quinn, who is currently in charge of the CERT tourism trade training body, was appointed by Tourism Minister John O'Donoghue.

****

1. Global poll shows increased anti-Americanism 2. Global survey on anti-Americanism A large survey of global public opinion has found evidence of increased anti-Americanism. However, it also revealed that most people thought the spread of science and technology from the United States was beneficial. The study by a Washington-based research committee, the Pew Global Attitudes Survey, involved 38,000 interviews in 44 countries and was headed by the former US Secretary of State, Madeleine Albright. The anti-American trend was most dramatic in muslim societies, such as Egypt and Pakistan. One of the research team, Bruce Stokes, said that although American culture appeared to be popular, US government policy was attracting growing criticism.

- **Identical**
- **Very similar**
- **Quite similar**
- **Quite different**
- **Very different**
- **Completely different**

(D.17) 1. New tourism body established 2. New tourism authority established The first head of a new tourism body established under the terms of the Good Friday agreement has been named. Shaun Quinn will be chief executive of the National Tourism Development Authority that will form part of the Tourism Ireland agency. He will take up the position in time for the start of the 2003 tourism season. The authority is to take over responsibility for developing the industry from Bord Failte, which will now concentrate on specialist functions, including sporting events such as the 2006 Ryder Cup golf. The NTDA will operate under the name Failte Ireland. Legislation to set it up was introduced last month in the Dail. Mr Quinn, who is currently in charge of the CERT tourism trade training body, was appointed by Tourism Minister John O'Donoghue.

****

1. Global poll shows increased anti-Americanism 2. Global survey on anti-Americanism A large survey of global public opinion has found evidence of increased anti-Americanism. However, it also revealed that most people thought the spread of science and technology from the United States was beneficial. The study by a Washington-based research committee, the Pew Global Attitudes Survey, involved 38,000 interviews in 44 countries and was headed by the former US Secretary of State, Madeleine Albright. The anti-American trend was most dramatic in muslim societies, such as Egypt and Pakistan. One of the research team, Bruce Stokes, said that although American culture appeared to be popular, US government policy was attracting growing criticism.

- Are these two stories the same?
- Are these two stories about the same event?
- Are these two stories about the same type of event?
- Are these two stories on the same topic?
- Are these two stories the same genre?
- Are these two stories similar?

(D.18) 1. New tourism body established 2. New tourism authority established The first head of a new tourism body established under the terms of the Good Friday agreement has been named. Shaun Quinn will be chief executive of the National Tourism Development Authority that will form part of the Tourism Ireland agency. He will take up the position in time for the start of the 2003 tourism season. The authority is to take over responsibility for developing the industry from Bord Failte, which will now concentrate on specialist functions, including sporting events such as the 2006 Ryder Cup golf. The NTDA will operate under the name Failte Ireland. Legislation to set it up was introduced last month in the Dail. Mr Quinn, who is currently in charge of the CERT tourism trade training body, was appointed by Tourism Minister John O'Donoghue.

****

1. Global poll shows increased anti-Americanism 2. Global survey on anti-Americanism A large survey of global public opinion has found evidence of increased anti-Americanism. However, it also revealed that most people thought the spread of science and technology from the United States was beneficial. The study by a Washington-based research committee, the Pew Global Attitudes Survey, involved 38,000 interviews in 44 countries and was headed by the former US Secretary of State, Madeleine Albright. The anti-American trend was most dramatic in muslim societies, such as Egypt and Pakistan. One of the research team, Bruce Stokes, said that although American culture appeared to be popular, US government policy was attracting growing criticism.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Identical	..	..	..	..	..	..	..	..	**Totally different**

(D.19) 1. Gardai begin Kilkenny murder hunt 2. Gardai begin Kilkenny murder hunt Gardai in Kilkenny have begun a murder investigation following the discovery of a woman's body in a house in the city. The 36-year-old mother of two was found by emergency services responding to reports of a fire at the house. At around 5.00pm yesterday evening the emergency services were called to a reported fire at a semi-detached house in Greenfield Road, Kilkenny City. The body of a woman was taken from the house and brought to St Luke's Hospital. It is understood the woman had extensive injuries to her head and torso. Gardai have now launched a full-scale murder investigation, and a post mortem examination is being carried out by the State Pathologist, Dr John Harbison. Members of the Garda Technical Bureau have begun examining the scene, which remains cordoned off. ********** 1. Man due in court for bomb attempt A 34-year-old man is due in court in the North charged in connection with an attempted firebomb attack in Belfast city centre at the weekend. The man has been charged with possession of explosives with intent to endanger life and conspiracy to cause an explosion. He was arrested in a major security operation in the city centre last Sunday night, during which a second man was shot twice by police. Police discovered a device consisting of two gas cylinders and two pipe bombs linked to a number of containers of flammable liquid in a car abandoned outside the motor tax office at Upper Queen Street at the weekend. A timer device had been attached and activated. The shooting of the second man by police is being investigated by the North's Police Ombudsman, Nuala O'Loan.

- **Identical**
- **Very similar**
- **Quite similar**
- **Quite different**
- **Very different**
- **Completely different**

(D.20) 1. Gardai begin Kilkenny murder hunt 2. Gardai begin Kilkenny murder hunt Gardai in Kilkenny have begun a murder investigation following the discovery of a woman's body in a house in the city. The 36-year-old mother of two was found by emergency services responding to reports of a fire at the house. At around 5.00pm yesterday evening the emergency services were called to a reported fire at a semi-detached house in Greenfield Road, Kilkenny City. The body of a woman was taken from the house and brought to St Luke's Hospital. It is understood the woman had extensive injuries to her head and torso. Gardai have now launched a full-scale murder investigation, and a post mortem examination is being carried out by the State Pathologist, Dr John Harbison. Members of the Garda Technical Bureau have begun examining the scene, which remains cordoned off. ********** 1. Man due in court for bomb attempt A 34-year-old man is due in court in the North charged in connection with an attempted firebomb attack in Belfast city centre at the weekend. The man has been charged with possession of explosives with intent to endanger life and conspiracy to cause an explosion. He was arrested in a major security operation in the city centre last Sunday night, during which a second man was shot twice by police. Police discovered a device consisting of two gas cylinders and two pipe bombs linked to a number of containers of flammable liquid in a car abandoned outside the motor tax office at Upper Queen Street at the weekend. A timer device had been attached and activated. The shooting of the second man by police is being investigated by the North's Police Ombudsman, Nuala O'Loan.

- Are these two stories the same?
- Are these two stories about the same event?
- Are these two stories about the same type of event?
- Are these two stories on the same topic?
- Are these two stories the same genre?
- Are these two stories similar?

(D.21) 1. Gardai begin Kilkenny murder hunt 2. Gardai begin Kilkenny murder hunt Gardai in Kilkenny have begun a murder investigation following the discovery of a woman's body in a house in the city. The 36-year-old mother of two was found by emergency services responding to reports of a fire at the house. At around 5.00pm yesterday evening the emergency services were called to a reported fire at a semi-detached house in Greenfield Road, Kilkenny City. The body of a woman was taken from the house and brought to St Luke's Hospital. It is understood the woman had extensive injuries to her head and torso. Gardai have now launched a full-scale murder investigation, and a post mortem examination is being carried out by the State Pathologist, Dr John Harbison. Members of the Garda Technical Bureau have begun examining the scene, which remains cordoned off. ********** 1. Man due in court for bomb attempt A 34-year-old man is due in court in the North charged in connection with an attempted firebomb attack in Belfast city centre at the weekend. The man has been charged with possession of explosives with intent to endanger life and conspiracy to cause an explosion. He was arrested in a major security operation in the city centre last Sunday night, during which a second man was shot twice by police. Police discovered a device consisting of two gas cylinders and two pipe bombs linked to a number of containers of flammable liquid in a car abandoned outside the motor tax office at Upper Queen Street at the weekend. A timer device had been attached and activated. The shooting of the second man by police is being investigated by the North's Police Ombudsman, Nuala O'Loan.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Identical	..	..	..	..	..	..	..	..	**Totally different**

(D.22) MORIARTY The Moriarty Tribunal has revealed a second incident where it appears that Denis O'Brien's consortium had access to confidential information relating to the competition for the second mobile phone. In July of 1995 Mr O'Brien's consortium got a copy of a European Commission letter sent to Michael Lowry. The letter gave confidential and valuable information regarding the importance that would be attached to the fee to be charged in the competition for the licence. Members of the consortium have claimed their copy of the letter came from the European Commission and they argue the information was not confidential but the tribunal disagrees. The tribunal says it will be examining whether the integrity of the competition process was compromised.
****

(credit Aoife Kavanagh) The case against the woman being sued over a Pyramid scheme has been dismissed by the Dublin District Court. Ms Lynda Laird was being sued by two investors who claimed Ms Laird had frustrated their efforts to recieve a promised 45, 000 euro in two months. Sister and brother Rose and Dermot Smedker said they invested 5,713 euro in the scheme. Delivering his judgement Judge O'Leary said they had accused Ms Laird of operating a scam but said if that was the case, the plantifs took part in the same scam. He said the scheme appeared to be a form of Pyramid investment but it fell outside the scope of the Pyramid Act. He said it was clear it worked on the basis that an early minority investing would win and a late majority would loose .He went on to rule against awarding any costs in this case.

- **Identical**
- **Very similar**
- **Quite similar**
- **Quite different**
- **Very different**
- **Completely different**

(D.23) MORIARTY The Moriarty Tribunal has revealed a second incident where it appears that Denis O'Brien's consortium had access to confidential information relating to the competition for the second mobile phone. In July of 1995 Mr O'Brien's consortium got a copy of a European Commission letter sent to Michael Lowry. The letter gave confidential and valuable information regarding the importance that would be attached to the fee to be charged in the competition for the licence. Members of the consortium have claimed their copy of the letter came from the European Commission and they argue the information was not confidential but the tribunal disagrees. The tribunal says it will be examining whether the integrity of the competition process was compromised.

****

(credit Aoife Kavanagh) The case against the woman being sued over a Pyramid scheme has been dismissed by the Dublin District Court. Ms Lynda Laird was being sued by two investors who claimed Ms Laird had frustrated their efforts to recieve a promised 45, 000 euro in two months. Sister and brother Rose and Dermot Smedker said they invested 5,713 euro in the scheme. Delivering his judgement Judge O'Leary said they had accused Ms Laird of operating a scam but said if that was the case, the plantifs took part in the same scam. He said the scheme appeared to be a form of Pyramid investment but it fell outside the scope of the Pyramid Act. He said it was clear it worked on the basis that an early minority investing would win and a late majority would loose .He went on to rule against awarding any costs in this case.

- Are these two stories the same?
- Are these two stories about the same event?
- Are these two stories about the same type of event?
- Are these two stories on the same topic?
- Are these two stories the same genre?
- Are these two stories similar?

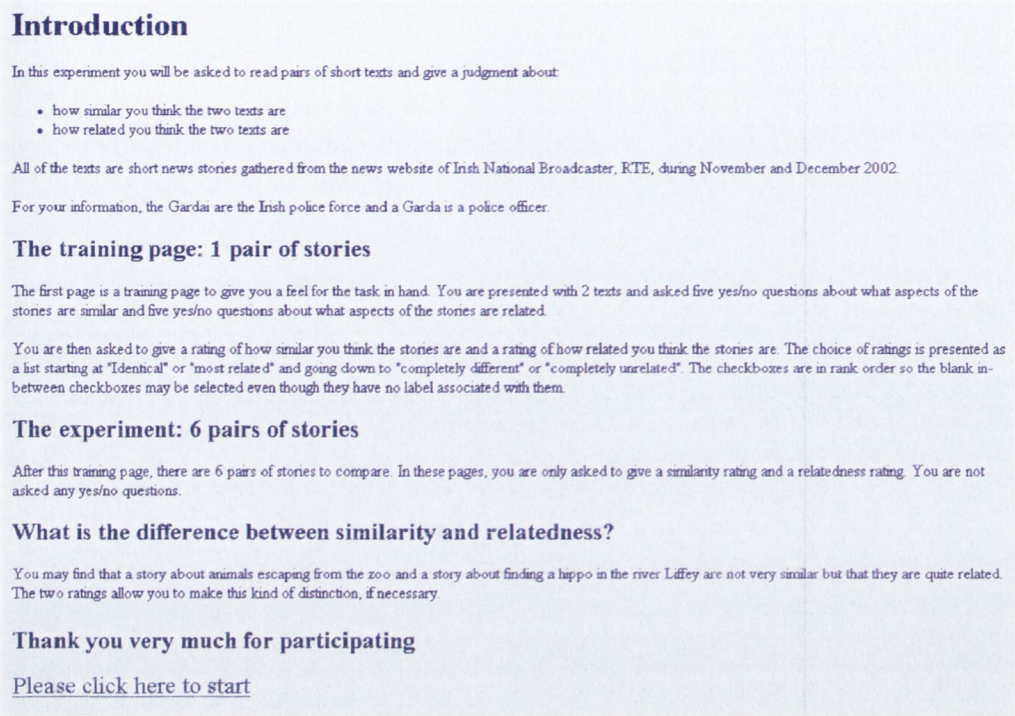
(D.24) MORIARTY The Moriarty Tribunal has revealed a second incident where it appears that Denis O'Brien's consortium had access to confidential information relating to the competition for the second mobile phone. In July of 1995 Mr O'Brien's consortium got a copy of a European Commission letter sent to Michael Lowry. The letter gave confidential and valuable information regarding the importance that would be attached to the fee to be charged in the competition for the licence. Members of the consortium have claimed their copy of the letter came from the European Commission and they argue the information was not confidential but the tribunal disagrees. The tribunal says it will be examining whether the integrity of the competition process was compromised.

****

(credit Aoife Kavanagh) The case against the woman being sued over a Pyramid scheme has been dismissed by the Dublin District Court. Ms Lynda Laird was being sued by two investors who claimed Ms Laird had frustrated their efforts to recieve a promised 45, 000 euro in two months. Sister and brother Rose and Dermot Smedker said they invested 5,713 euro in the scheme. Delivering his judgement Judge O'Leary said they had accused Ms Laird of operating a scam but said if that was the case, the plantifs took part in the same scam. He said the scheme appeared to be a form of Pyramid investment but it fell outside the scope of the Pyramid Act. He said it was clear it worked on the basis that an early minority investing would win and a late majority would loose .He went on to rule against awarding any costs in this case.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Identical	..	..	..	..	..	..	..	..	**Totally different**

D.2 Gold Standard Experiment Screen Shots



Introduction

In this experiment you will be asked to read pairs of short texts and give a judgment about:

- how similar you think the two texts are
- how related you think the two texts are

All of the texts are short news stories gathered from the news website of Irish National Broadcaster, RTE, during November and December 2002.

For your information, the Gardai are the Irish police force and a Garda is a police officer.

The training page: 1 pair of stories

The first page is a training page to give you a feel for the task in hand. You are presented with 2 texts and asked five yes/no questions about what aspects of the stories are similar and five yes/no questions about what aspects of the stories are related.

You are then asked to give a rating of how similar you think the stories are and a rating of how related you think the stories are. The choice of ratings is presented as a list starting at "Identical" or "most related" and going down to "completely different" or "completely unrelated". The checkboxes are in rank order so the blank in-between checkboxes may be selected even though they have no label associated with them.

The experiment: 6 pairs of stories

After this training page, there are 6 pairs of stories to compare. In these pages, you are only asked to give a similarity rating and a relatedness rating. You are not asked any yes/no questions.

What is the difference between similarity and relatedness?

You may find that a story about animals escaping from the zoo and a story about finding a hippo in the river Liffey are not very similar but that they are quite related. The two ratings allow you to make this kind of distinction, if necessary.

Thank you very much for participating

Please click here to start

Figure D.1: Experiment Instructions Page Screenshot

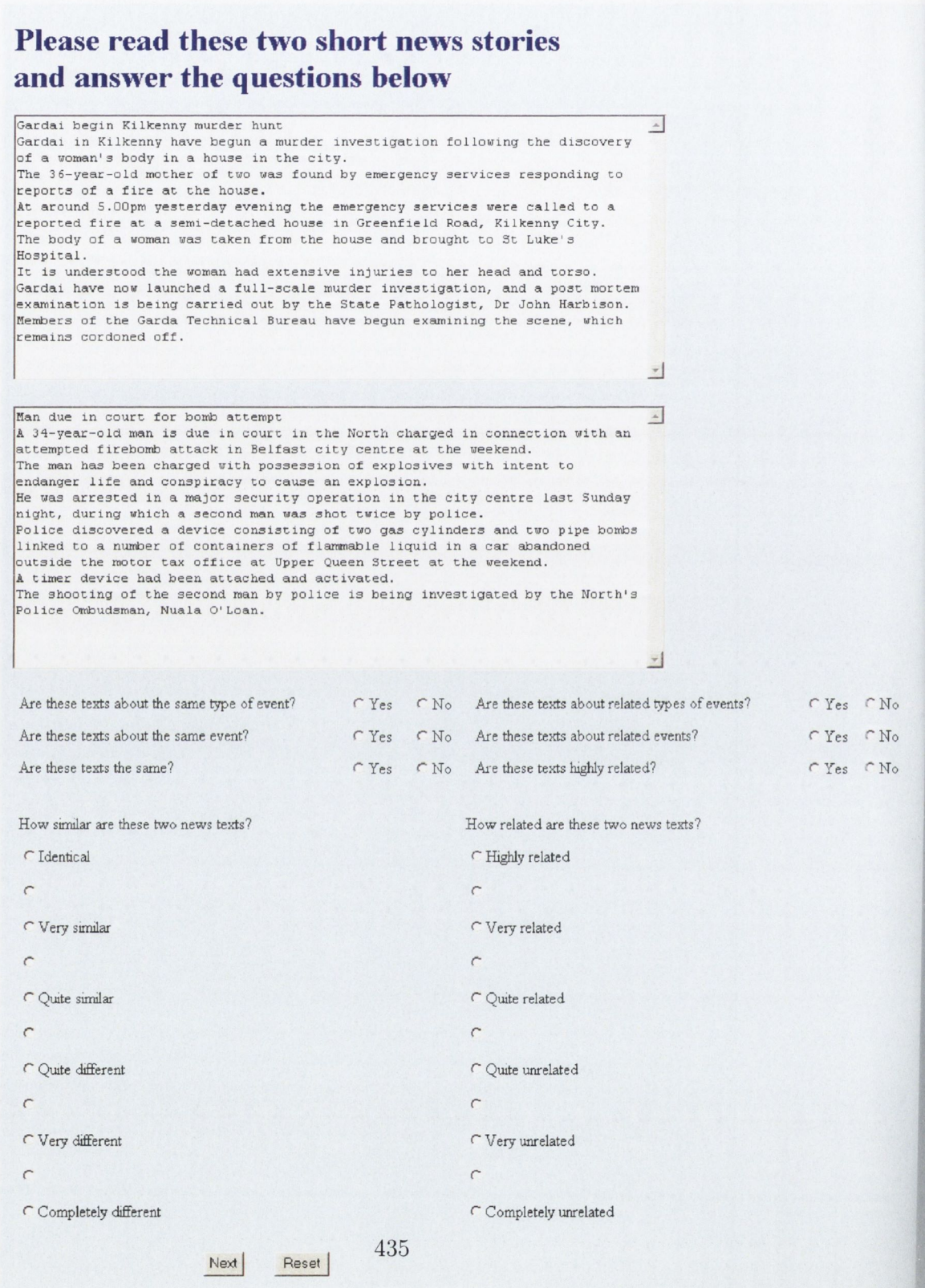


Figure D.2: Experiment Training Page Screenshot

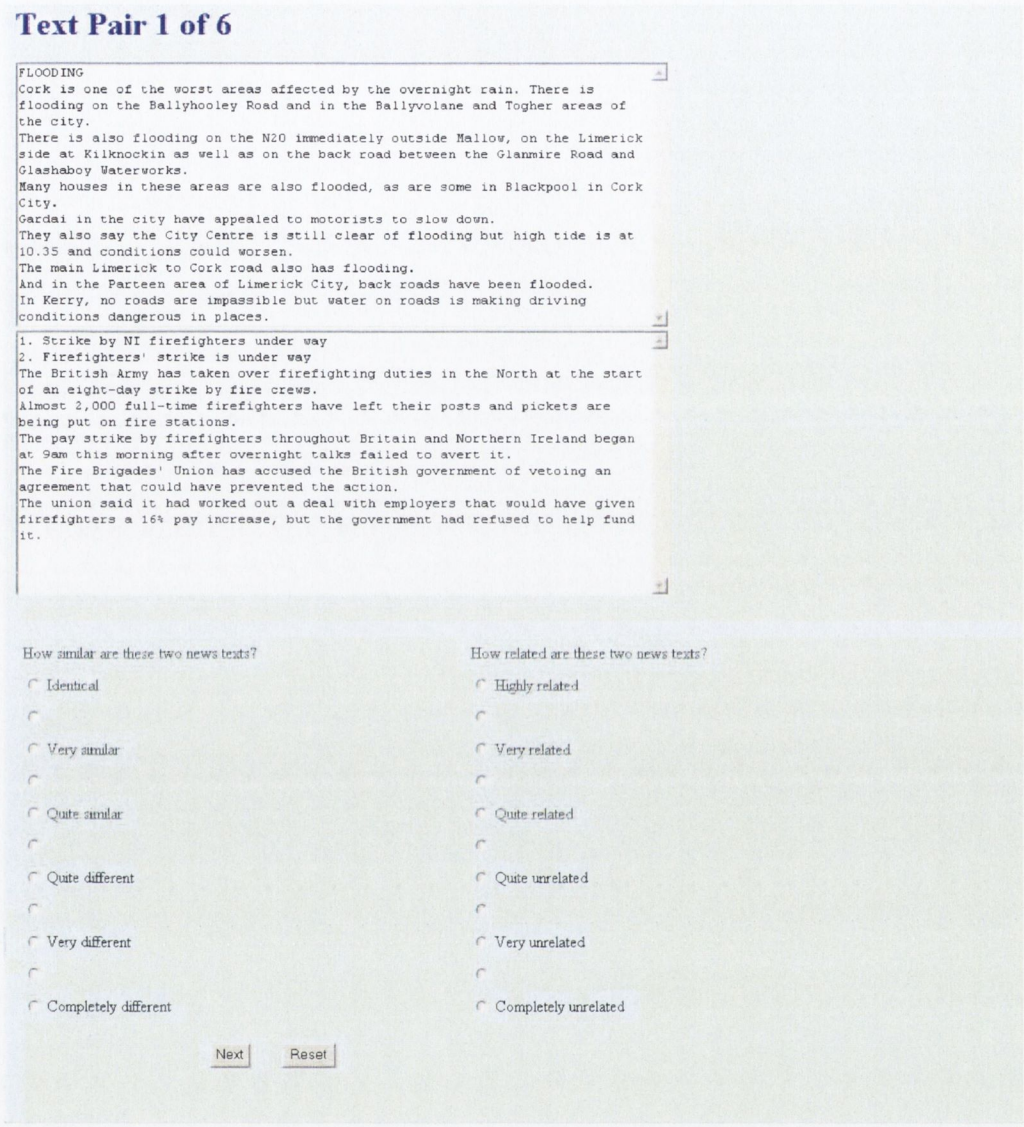


Figure D.3: Sample Experiment Page Screenshot

D.3 Gold Standard Experiment Source Code

```
#!/usr/local/bin/perl -w
# hiweb - load CGI module

################################################
# SimilarityExperiment -- a script to run an experiment on-line to compare texts
# Outputs errors to mycgi-log
# Outputs experiment results to data.out

# Compile time things
# Import CGI
# Redirect STDERR to a log file
################################################

BEGIN {
    use strict;
    use Fcntl qw(:DEFAULT :flock);
    use IO::Seekable;
    use CGI::Carp qw(carpout);
    use CGI qw(:standard escapeHTML);
    open(LOG, ">>.mycgi-log") or die "Unable to append to mycgi-log: $!\n";
    carpout(*LOG);
}


################################################
# HTML blurb used in the program.
# Customised for colour etc
################################################
my $noOfTextsPerPerson=6;
my %Texts = (
  '271102_AMDIraq8' => \&func271102_AMDIraq8,
  '041202TM1745_AMDBudgetSources25' => \&func041202TM1745_AMDBudgetSources25,
  '041202TM1745_AMDBudgetSources14' => \&func041202TM1745_AMDBudgetSources14,
  '061202_AMDMurder4' => \&func061202_AMDMurder4,
  '041202TM1513_AMDPyramid2' => \&func041202TM1513_AMDPyramid2,
  '271102_AMDFloods6' => \&func271102_AMDFloods6,
  '041202TM1745_AMDBudgetReaction5' => \&func041202TM1745_AMDBudgetReaction5,
  '221102_AMDflood9' => \&func221102_AMDflood9,
  '041202TM1513_AMDEclipse3' => \&func041202TM1513_AMDEclipse3,
  '041202TM1745_AMDMoriarty3' => \&func041202TM1745_AMDMoriarty3,
  '051202_AMDMoriarty2' => \&func051202_AMDMoriarty2,
  '041202TM1537_AMDTourism2' => \&func041202TM1537_AMDTourism2,
  '041202TM1104AMDIraq5' => \&func041202TM1104AMDIraq5,
  '051202_AMDAnti-US3' => \&func051202_AMDAntiUS3,
  '041202TM1745_AMDBudgetSources49' => \&func041202TM1745_AMDBudgetSources49,
  '271102_AMDIsrael8' => \&func271102_AMDIsrael8,
  '051202_AMDProdi7' => \&func051202_AMDProdi7,
  '061202_AMDMacDoBomb2' => \&func061202_AMDMacDoBomb2,
  '221102_AMDpisces5' => \&func221102_AMDpisces5,
  '271102_AMDFloods4' => \&func271102_AMDFloods4,
  '041202TM1745_AMDBudgetSources17' => \&func041202TM1745_AMDBudgetSources17,
  '221102_AMDflood8' => \&func221102_AMDflood8,
  '221102_AMDfirefighters8' => \&func221102_AMDfirefighters8,
  '051202_AMDFlood2' => \&func051202_AMDFlood2,
```

```
  '271102AMDFirebomb4' => \&func271102AMDFirebomb4
  );

my $startBlurb="<?xml version=\"1.0\" encoding=\"iso-8859-1\"?>\n<!DOCTYPE html\nPUBLIC \"-//W3C//DTD XHTML 1.0 Transitional//EN\"\n
\"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd\">\n<html xmlns=\"http://www.w3.org/1999/xhtml\" lang=\"en-US\"><head>
<title>Similarity Judgment Experiment</title>\n
<SCRIPT LANGUAGE=\"JavaScript\">\nfunction checkFields(){\n    var simChecked=false\n
var checkedFlag=true;\n    for(i=0;i<document.FORM1.simRating.length;i++){\n
if(document.FORM1.simRating[i].checked==true){\n  simChecked=true\n  break\n        }\n    }\n
if(simChecked==false){\n checkedFlag=false\n alert(\"Please click a similarity rating\")\n        return checkedFlag\n    }\n
return checkedFlag\n}\n</SCRIPT>\n</head><body BGCOLOR=\"#ccffff\">\n<font color=\"darkblue\">\n";

my $startTable="<TABLE BORDER=\"0\" BGCOLOR=\"#ccffff\" WIDTH=\"90%\" cellpadding=\"5\" rules=\"groups\">\n
<colgroup width=\"50%\" align=\"left\">\n<colgroup width=\"50%\" align=\"left\">\n
<TR><TD><LABEL FOR=\"simRating\">How similar are these two news texts?</LABEL></TD>\n
<TD><LABEL FOR=\"relRating\">How related are these two news texts?</LABEL></TD></TR>\n";

my $endTable="</TABLE>\n";
my $startButtonTable="<TABLE BORDER=\"0\" WIDTH=\"50%\" cellspacing=\"10\" cellpadding=\"10\">\n<TR>\n";
my $submitButton="<TD WIDTH=\"25%\" ALIGN=\"right\">";
my $resetButton="<TD WIDTH=\"25%\" ALIGN=\"left\">";
my $endButton="</TD>";
my $endButtonTable="</TR>\n</TABLE>";

################################################
# Some form constants
# method, action and radiobox labels
################################################
my $method="POST";
my $onSubmit="return checkFields()";
my $action="http://www.cs.tcd.ie/Ann.Devitt/cgi-bin/Experiment.cgi";
my @simLabels=("  Identical", " ", "  Very similar", " ",
"  Quite similar", " ", "  Quite different", " ", "  Very different", " ", "  Completely different");
my @relLabels=("  Highly related", " ", "  Very related", " ", "
Quite related", " ", "  Quite unrelated", " ", "  Very unrelated", " ", "  Completely unrelated");
my $remote_host=remote_host();
my $now_string=localtime;
my $counter=param("counter");
my $user=-1;
################################################
# Handling the output
# Open the output file
# Get remote host and timestamp
# Ignore text1 and text2 values (i.e. the actual stories)
# Print remote host, timestamp and parameters to the output file
################################################

if($counter==0){
    my $q = new CGI;
    $user=&getUserNumber;
    $filename="Results/exp$user.out";
    open (OUT,">>$filename") || die "Can't open output file $filename: $!\n";
    print OUT "remoteHost=$remote_host\n";
    print OUT "timestamp=$now_string\n";
    print OUT "user=$user\n";
    $q->param(-name=>'counter',-value=>$_);
```

```
    $q->param(-name=>'text1',-value=>"");
    $q->param(-name=>'text2',-value=>"");
    $q->save(OUT);
}
else {
    $textpair=param("textpair");
    $user=param("user");
    $filename="Results/exp$user.out";
    open (OUT,">>$filename") || die "Can't open output file $filename: $!\n";
    $simRating=param("simRating");
    $relRating=param("relRating");
    if(!(defined($relRating))){
$relRating=-1;
    }
    print OUT "remoteHost=$remote_host\n";
    print OUT "timestamp=$now_string\n";
    print OUT "textpair=$textpair\n";
    print OUT "counter=$counter\n";
    print OUT "user=$user\n";
    print OUT "simRating=$simRating\n";
    print OUT "relRating=$relRating\n";
    print OUT "=\n";
}
close OUT;

################################################
# Generate the next form
# The form contains
# textarea1: text1
# textarea2: text2
# textpair: hidden field containing the names of the files
# counter: hidden field containing how many files are done
# simRating: table of radioboxes for similarity judgement
# relRating: table of radioboxes for relatedness judgement
################################################

if($counter<$noOfTextsPerPerson){
################################################
## get texts will be using
################################################
    my $errorOccurred=0;
    $counter=$counter+1;
    $currentCombo=&getCurrentTextCombo($counter, $user);
    @currentText=split(":", $currentCombo);
    ## print out a new file ##
    print header();
    print $startBlurb;
    print "<H1>Text Pair $counter of 6</H1>\n";
    print start_form(-method=>$method, -action=>$action, -name=>"FORM1", -onSubmit=>$onSubmit);
    $first=int(rand(2));
    $second=($first+1)%2;
    if($Texts{$currentText[$first]}){
$Texts{$currentText[$first]}->();
    }
    else {
$errorOccurred=1;
```

```
&funcErrorMessage();
print STDERR "Error printing textarea1 for user $user and counter $counter\n";
print STDERR "index: $first\tfilename:\t$currentText[$first]\tsuroutine call:$Texts{$currentText[$first]}";
    }
    if($Texts{$currentText[$second]}){
$Texts{$currentText[$second]}->();
    }
    else {
$errorOccurred=1;
&funcErrorMessage();
print STDERR "Error printing textarea2 for user $user and counter $counter\n";
print STDERR "index: $second\tfilename:\t$currentText[$second]\tsubroutine call:$Texts{$currentText[$second]}";
    }
    print p();
    if($errorOccurred==1){
print "<INPUT TYPE=\"HIDDEN\" NAME=\"textpair\" VALUE=\"Error:$currentCombo\"/>";
    }
    else {
print "<INPUT TYPE=\"HIDDEN\" NAME=\"textpair\" VALUE=\"$currentText[0]:$currentText[1]\"/>";
    }
    print "<INPUT TYPE=\"HIDDEN\" NAME=\"counter\" VALUE=\"$counter\"/>";
    print "<INPUT TYPE=\"HIDDEN\" NAME=\"user\" VALUE=\"$user\"/>";
    print "$startTable";
    for $i (0 .. $#simLabels){
print "<TR>";
print "<TD><INPUT TYPE=\"RADIO\" NAME=\"simRating\" ID=\"$i\" VALUE=\"$i\">$simLabels[$i]</TD>";
print "<TD><INPUT TYPE=\"RADIO\" NAME=\"relRating\" ID=\"$i\" VALUE=\"$i\">$relLabels[$i]</TD></TR>";
    }
    print "$endTable";
    print "$startButtonTable";
    print "$submitButton";
    if($counter==$noOfTextsPerPerson){
print submit(-name=>"Submit", -value=>"Finish");
    }
    else {
print submit(-name=>"Next", -value=>"Next");
    }
    print "$endButton";
    print "$resetButton";
    print reset();
    print "$endButton";
    print "$endButtonTable";
    print endform();
    print end_html();
}

################################################
# Final comments to the user
################################################

else {
    print header();
    print "$startBlurb\n";
    print h1("Thank you very much");
    print h1("If you have any queries please contact me at devitta at cs.tcd.ie");
    print end_html();
```

```
}


## SUBROUTINES
sub getUserNumber {
    my $sizeof=4;
    my $buffer;
    my $answer;
    my $newnumber;
    open(FILE, "+< .Users.dat") or die "can't open this ";
    flock(FILE, LOCK_EX) or die "can't lock file $!";
    seek(FILE, 0, 0) or die "seek failed: $!";
    read(FILE, $buffer, 4)==$sizeof or die "read failed: $!";
    if($buffer=~m/^0*([1-9])([0-9]*)$/){
$answer=$1.$2;
$newnumber=$answer+1;
$length=length($newnumber);
for($i=$length;$i<$sizeof;$i++){
    $newnumber="0". $newnumber;
}
    }
    seek(FILE, 0, 0) or die "seek 2 failed: $!";
    print FILE "$newnumber\n";
    close(FILE) or die "closed failed";
    return $answer;
}

sub getCurrentTextCombo {
    my ($count, $userNumber)=@_;
    $userNumber--;
    if($userNumber>499){
$userNumber=$userNumber%500;
    }
    my $comboFile=".FileCombos.dat";
    my $comboIndex=".FileCombos.index";
    open(COMBOS, "< $comboFile") or die "Couldn't open file: $comboFile: $!\n";
    open(IDX, "< $comboIndex") or die "Can't open index $comboIndex for read: $!";
    my $linenumber=($userNumber*6)+$count;
    my $size;
    my $i_offset;
    my $entry;
    my $d_offset;
    $size=length(pack("N",0));
    $i_offset=$size*($linenumber-1);
    seek(IDX, $i_offset, 0) or return;
    read(IDX, $entry, $size);
    $d_offset=unpack("N", $entry);
    seek(COMBOS, $d_offset, 0);
    $line= scalar(<COMBOS>);
    chomp($line);
    close(IDX);
    close(COMBOS);
    return $line;
}

sub funcErrorMessage{
```

```
    print "<textarea name=\"text1\" id=\"text1\" rows=16 cols=78 wrap=\"virtual\" readonly=\"true\">\n";
    print "An error occurred\n";
    print "Please continue to next slide\n";
    print "</textarea>";
}

sub func271102_AMDIraq8{
    print "<textarea name=\"text1\" id=\"text1\" rows=16 cols=78 wrap=\"virtual\" readonly=\"true\">\n";
    print "1. UN weapons inspectors return to Iraq\n";
    print "2. UN arms inspectors return to Iraq\n";
    print "United Nations weapons inspectors have resumed inspections in Iraq under the terms of a
           strongly worded Security Council resolution. \n";
    print "A team of seventeen experts have carried out the first checks on suspected arms sites in nearly four years.
          They will be joined by a further 300 inspectors over the coming months.\n";
    print "The UN Secretary General, Kofi Annan, has underlined that military action against Iraq is avoidable if
          President Saddam Hussein honours his commitments and co-operates fully with the inspectors.\n";
    print "Mr Annan said that if Iraq acted as it should, the arguments in favour of war would be considerably weakened.\n";
    print "</textarea>";
}

## The 24 other sub routines to print each story were omitted
## The format is the same as subroutine func271102_AMDIraq8 given above
```

Appendix E

Gold Standard Experiment Results

E.1 Group Rating Tables

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	8	8	10	10	10	6	8.67
User2	-	-	-	-	-	-	
User3	10	10	10	10	10	9	9.83
User4	10	9	-	-	-	-	9.50
User5	-	-	-	-	-	-	
User6	10	10	10	10	9	2	8.50
User7	8	8	7	8	6	6	7.17
User8	10	10	10	10	8	8	9.33
User9	10	6	10	8	6	2	7.00
User10	-	-	-	-	-	-	
Average	9.43	8.71	9.50	9.33	8.17	5.50	

Table E.1: Group 1 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	10	10	10	10	10	10	10.00
User2	10	10	10	10	10	8	9.67
User3	10	8	8	10	10	10	9.33
User4	9	7	6	8	8	8	7.67
User5	10	10	10	10	10	10	10.00
User6	10	10	10	10	10	10	10.00
User7	10	10	8	10	10	10	9.67
User8	10	8	8	10	10	10	9.33
User9	10	10	10	10	10	10	10.00
User10	8	6	8	6	6	10	7.33
Average	9.70	8.90	8.80	9.40	9.40	9.60	

Table E.2: Group 2 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	8	10	10	10	10	10	9.67
User2	8	10	10	8	10	10	9.33
User3	8	10	10	5	10	10	8.83
User4	8	8	-	-	-	-	8.00
User5	6	10	10	8	10	10	9.00
User6	2	10	10	8	8	10	8.00
User7	3	7	5	6	5	8	5.67
User8	10	10	10	8	8	10	9.33
User9	10	10	10	10	10	10	10.00
User10	10	10	10	10	9	10	9.83
Average	7.30	9.50	9.44	8.11	8.89	9.78	

Table E.3: Group 3 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	-	-	-	-	-	-	
User2	8	10	10	2	8	10	8.00
User3	10	10	8	10	8	10	9.33
User4	10	4	10	9	8	10	8.50
User5	10	10	10	8	10	10	9.67
User6	10	10	10	10	10	10	10.00
User7	10	4	5	5	7	5	6.00
User8	6	10	10	10	10	8	9.00
User9	10	8	6	8	10	8	8.33
User10	-	-	-	-	-	-	
Average	9.25	8.25	8.62	7.75	8.88	8.88	

Table E.4: Group 4 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	10	10	10	7	8	7	8.67
User2	6	5	8	6	5	5	5.83
User3	8	8	4	8	7	9	7.33
User4	10	10	6	10	10	10	9.33
User5	-	-	-	-	-	-	
User6	10	10	8	10	10	10	9.67
User7	6	4	6	6	4	4	5.00
User8	6	8	8	8	8	10	8.00
User9	10	7	8	5	9	8	7.83
User10	8	-	-	-	-	-	8.00
Average	8.22	7.75	7.25	7.50	7.62	7.88	

Table E.5: Group 5 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	10	10	8	10	8	10	9.33
User2	8	10	8	8	10	8	8.67
User3	6	2	6	2	8	8	5.33
User4	4	6	6	6	4	8	5.67
User5	-	-	-	-	-	-	
User6	8	6	10	8	4	8	7.33
User7	-	-	-	-	-	-	
User8	8	6	8	10	4	8	7.33
User9	-	-	-	-	-	-	
User10	-	-	-	-	-	-	
Average	7.33	6.67	7.67	7.33	6.33	8.33	

Table E.6: Group 6 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	10	10	10	4	10	9	8.83
User2	7	8	6	6	6	8	6.83
User3	8	8	6	8	-	-	7.50
User4	10	8	6	5	8	8	7.50
User5	10	10	10	8	10	8	9.33
User6	4	8	-	-	-	-	6.00
User7	10	10	10	8	10	10	9.67
User8	8	6	4	6	8	4	6.00
User9	10	10	10	10	10	10	10.00
User10	-	-	-	-	-	-	
Average	8.56	8.67	7.75	6.88	8.86	8.14	

Table E.7: Group 7 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	10	6	8	8	8	8	8.00
User2	10	6	7	8	8	8	7.83
User3	10	6	4	8	8	10	7.67
User4	10	10	10	10	10	10	10.00
User5	10	10	8	9	8	10	9.17
User6	8	6	10	10	9	10	8.83
User7	-	-	-	-	-	-	
User8	10	10	10	10	8	10	9.67
User9	10	7	8	9	6	8	8.00
User10	8	5	8	8	10	8	7.83
Average	9.56	7.33	8.11	8.89	8.33	9.11	

Table E.8: Group 8 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	4	4	4	4	3	8	4.50
User2	10	8	5	6	8	10	7.83
User3	10	8	8	10	4	10	8.33
User4	8	7	8	8	4	10	7.50
User5	10	10	10	10	10	10	10.00
User6	10	7	10	10	5	10	8.67
User7	10	6	6	6	4	8	6.67
User8	10	6	8	10	8	10	8.67
User9	6	8	6	6	2	10	6.33
User10	-	-	-	-	-	-	
Average	8.67	7.11	7.22	7.78	5.33	9.56	

Table E.9: Group 9 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	8	8	8	8	8	8	8.00
User2	-	-	-	-	-	-	
User3	8	6	6	6	6	6	6.33
User4	-	-	-	-	-	-	
User5	7	8	8	9	5	7	7.33
User6	10	10	10	10	8	10	9.67
User7	8	9	8	9	8	6	8.00
User8	10	10	10	10	8	10	9.67
User9	5	5	4	5	5	6	5.00
User10	10	8	8	10	4	8	8.00
Average	8.25	8.00	7.75	8.38	6.50	7.62	

Table E.10: Group 10 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	-	-	-	-	-	-	
User2	10	10	2	8	8	10	8.00
User3	10	10	4	6	6	4	6.67
User4	10	-	-	-	-	-	10.00
User5	5	5	2	5	5	8	5.00
User6	10	10	7	10	10	10	9.50
User7	10	10	6	10	10	10	9.33
User8	8	8	6	6	7	7	7.00
User9	10	10	8	10	10	10	9.67
User10	10	6	2	6	10	10	7.33
Average	9.22	8.62	4.62	7.62	8.25	8.62	

Table E.11: Group 11 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	8	-	-	-	-	-	8.00
User2	5	4	6	4	6	4	4.83
User3	10	10	4	10	6	6	7.67
User4	6	8	6	10	8	4	7.00
User5	8	10	4	10	8	4	7.33
User6	-	-	-	-	-	-	
User7	10	10	8	10	10	9	9.50
User8	8	6	4	10	6	2	6.00
User9	7	8	5	10	10	5	7.50
User10	7	8	4	6	7	3	5.83
Average	7.67	8.00	5.12	8.75	7.62	4.62	

Table E.12: Group 12 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	5	10	6	8	8	10	7.83
User2	1	10	10	10	10	10	8.50
User3	10	10	10	10	10	10	10.00
User4	6	10	6	6	6	10	7.33
User5	10	4	10	10	10	10	9.00
User6	6	6	8	8	6	8	7.00
User7	4	-	-	-	-	-	4.00
User8	6	6	6	6	8	8	6.67
User9	4	10	10	10	10	10	9.00
User10	5	10	10	10	10	10	9.17
Average	5.70	8.44	8.44	8.67	8.67	9.56	

Table E.13: Group 13 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	10	10	10	10	10	10	10.00
User2	8	10	10	8	10	8	9.00
User3	8	10	10	10	8	10	9.33
User4	10	8	8	8	10	10	9.00
User5	10	10	10	10	10	10	10.00
User6	8	4	10	6	10	8	7.67
User7	-	-	-	-	-	-	
User8	8	10	10	10	10	10	9.67
User9	10	10	10	10	10	10	10.00
User10	8	8	8	8	8	8	8.00
Average	8.89	8.89	9.56	8.89	9.56	9.33	

Table E.14: Group 14 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	6	3	10	10	10	10	8.17
User2	10	-	-	-	-	-	10.00
User3	10	10	10	10	10	10	10.00
User4	8	3	10	10	10	10	8.50
User5	10	6	-	-	-	-	8.00
User6	10	10	10	10	10	10	10.00
User7	6	4	4	6	4	3	4.50
User8	-	-	-	-	-	-	
User9	10	8	10	10	10	10	9.67
User10	10	4	9	10	9	10	8.67
Average	8.89	6.00	9.00	9.43	9.00	9.00	

Table E.15: Group 15 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	3	8	8	5	4	6	5.67
User2	-	-	-	-	-	-	
User3	8	9	10	10	7	10	9.00
User4	4	6	6	6	8	10	6.67
User5	3	5	2	4	5	5	4.00
User6	-	-	-	-	-	-	
User7	-	-	-	-	-	-	
User8	6	10	10	10	10	10	9.33
User9	8	8	10	10	10	10	9.33
User10	10	10	10	10	10	10	10.00
Average	6.00	8.00	8.00	7.86	7.71	8.71	

Table E.16: Group 16 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	4	8	2	6	10	4	5.67
User2	8	8	2	8	10	8	7.33
User3	10	10	3	6	10	8	7.83
User4	-	-	-	-	-	-	
User5	10	10	6	10	10	2	8.00
User6	-	-	-	-	-	-	
User7	6	6	6	8	8	8	7.00
User8	10	8	4	6	8	8	7.33
User9	8	8	4	6	8	6	6.67
User10	-	-	-	-	-	-	
Average	8.00	8.29	3.86	7.14	9.14	6.29	

Table E.17: Group 17 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	10	8	-	-	-	-	9.00
User2	8	8	8	8	6	6	7.33
User3	10	8	6	4	8	8	7.33
User4	8	6	6	4	8	8	6.67
User5	10	7	8	2	9	10	7.67
User6	10	10	10	4	10	10	9.00
User7	10	10	10	4	10	10	9.00
User8	6	8	8	6	8	8	7.33
User9	8	6	8	6	6	6	6.67
User10	-	-	-	-	-	-	
Average	8.89	7.89	8.00	4.75	8.12	8.25	

Table E.18: Group 18 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	2	4	10	9	8	4	6.17
User2	2	10	10	10	10	10	8.67
User3	-	-	-	-	-	-	
User4	-	-	-	-	-	-	
User5	-	-	-	-	-	-	
User6	8	8	8	8	8	8	8.00
User7	6	10	10	10	10	10	9.33
User8	-	-	-	-	-	-	
User9	10	10	10	10	10	10	10.00
User10	-	-	-	-	-	-	
Average	5.60	8.40	9.60	9.40	9.20	8.40	

Table E.19: Group 19 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	10	10	10	10	10	10	10.00
User2	10	10	10	10	10	10	10.00
User3	10	10	10	10	10	8	9.67
User4	10	10	10	10	10	10	10.00
User5	8	4	4	2	6	4	4.67
User6	9	6	-	-	-	-	7.50
User7	10	10	10	10	10	10	10.00
User8	-	-	-	-	-	-	
User9	10	8	10	6	-	-	8.50
User10	10	10	10	10	10	10	10.00
Average	9.67	8.67	9.25	8.50	9.43	8.86	

Table E.20: Group 20 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	7	6	7	8	6	9	7.17
User2	8	8	8	8	8	8	8.00
User3	4	10	10	10	10	10	9.00
User4	10	8	6	6	4	6	6.67
User5	10	10	10	10	10	10	10.00
User6	8	8	8	10	10	8	8.67
User7	6	8	8	8	6	8	7.33
User8	10	10	10	10	10	10	10.00
User9	6	8	8	10	8	8	8.00
User10	8	4	10	10	8	10	8.33
Average	7.70	8.00	8.50	9.00	8.00	8.70	

Table E.21: Group 21 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	8	8	2	10	10	8	7.67
User2	-	-	-	-	-	-	
User3	-	-	-	-	-	-	
User4	-	-	-	-	-	-	
User5	10	10	2	10	10	8	8.33
User6	10	10	5	10	10	10	9.17
User7	6	4	2	8	8	8	6.00
User8	-	-	-	-	-	-	
User9	4	6	2	8	6	4	5.00
User10	-	-	-	-	-	-	
Average	7.60	7.60	2.60	9.20	8.80	7.60	

Table E.22: Group 22 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	10	4	-	-	-	-	7.00
User2	-	-	-	-	-	-	
User3	6	6	4	6	6	6	5.67
User4	-	-	-	-	-	-	
User5	10	8	8	10	10	10	9.33
User6	6	5	2	5	6	5	4.83
User7	6	6	5	10	9	8	7.33
User8	4	-	-	-	-	-	4.00
User9	2	2	2	6	2	6	3.33
User10	4	3	5	6	6	6	5.00
Average	6.00	4.86	4.33	7.17	6.50	6.83	

Table E.23: Group 23 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	4	4	4	4	4	8	4.67
User2	6	10	9	5	8	8	7.67
User3	10	10	10	10	10	10	10.00
User4	10	6	8	10	10	10	9.00
User5	10	6	4	6	10	8	7.33
User6	6	8	8	4	10	10	7.67
User7	-	-	-	-	-	-	
User8	8	8	10	5	10	8	8.17
User9	-	-	-	-	-	-	
User10	-	-	-	-	-	-	
Average	7.71	7.43	7.57	6.29	8.86	8.86	

Table E.24: Group 24 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	8	8	-	-	-	-	8.00
User2	10	8	0	2	6	10	6.00
User3	10	10	2	4	6	8	6.67
User4	10	8	6	6	8	10	8.00
User5	8	6	5	2	4	6	5.17
User6	10	4	4	8	4	2	5.33
User7	10	-	-	-	-	-	10.00
User8	10	8	2	2	8	10	6.67
User9	10	6	2	8	8	8	7.00
User10	4	8	4	4	6	6	5.33
Average	9.00	7.33	3.12	4.50	6.25	7.50	

Table E.25: Group 25 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	2	6	6	10	10	4	6.33
User2	-	-	-	-	-	-	
User3	10	10	10	10	10	10	10.00
User4	10	10	-	-	-	-	10.00
User5	5	5	2	10	10	10	7.00
User6	-	-	-	-	-	-	
User7	10	2	4	10	2	6	5.67
User8	10	8	10	10	10	10	9.67
User9	-	-	-	-	-	-	
User10	10	10	8	10	10	10	9.67
Average	8.14	7.29	6.67	10.00	8.67	8.33	

Table E.26: Group 26 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	3	6	8	10	8	10	7.50
User2	8	8	10	10	4	10	8.33
User3	-	-	-	-	-	-	
User4	6	6	8	8	10	2	6.67
User5	8	8	10	10	10	10	9.33
User6	6	4	-	-	-	-	5.00
User7	4	10	10	10	10	10	9.00
User8	8	8	10	10	10	10	9.33
User9	-	-	-	-	-	-	
User10	10	10	10	10	8	10	9.67
Average	6.62	7.50	9.43	9.71	8.57	8.86	

Table E.27: Group 27 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	6	7	4	7	10	9	7.17
User2	6	10	10	4	8	4	7.00
User3	8	-	-	-	-	-	8.00
User4	8	6	-	-	-	-	7.00
User5	10	10	10	10	10	10	10.00
User6	10	10	6	6	8	10	8.33
User7	10	10	10	6	6	10	8.67
User8	10	10	8	6	6	10	8.33
User9	10	10	8	4	4	10	7.67
User10	-	-	-	-	-	-	
Average	8.67	9.12	8.00	6.14	7.43	9.00	

Table E.28: Group 28 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	-	-	-	-	-	-	
User2	-	-	-	-	-	-	
User3	-	-	-	-	-	-	
User4	10	2	10	4	2	10	6.33
User5	5	5	7	5	8	8	6.33
User6	-	-	-	-	-	-	
User7	10	10	10	6	10	10	9.33
User8	6	10	10	10	10	10	9.33
User9	8	10	10	6	6	8	8.00
User10	8	10	8	10	8	10	9.00
Average	7.83	7.83	9.17	6.83	7.33	9.33	

Table E.29: Group 29 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	6	10	6	10	6	8	7.67
User2	6	10	4	8	6	6	6.67
User3	-	-	-	-	-	-	
User4	9	10	6	9	10	10	9.00
User5	-	-	-	-	-	-	
User6	10	-	-	-	-	-	10.00
User7	10	10	9	10	5	10	9.00
User8	10	10	6	10	10	6	8.67
User9	10	10	6	10	10	6	8.67
User10	4	10	8	10	6	10	8.00
Average	8.12	10.00	6.43	9.57	7.57	8.00	

Table E.30: Group 30 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	8	4	4	6	4	8	5.67
User2	-	-	-	-	-	-	
User3	10	10	10	10	10	10	10.00
User4	10	4	4	2	10	4	5.67
User5	-	-	-	-	-	-	
User6	10	10	10	10	10	10	10.00
User7	10	10	10	10	10	10	10.00
User8	10	-	-	-	-	-	10.00
User9	10	10	8	6	8	6	8.00
User10	10	5	10	6	9	8	8.00
Average	9.75	7.57	8.00	7.14	8.71	8.00	

Table E.31: Group 31 table of similarity ratings

User	Pair1	Pair2	Pair3	Pair4	Pair5	Pair6	User Average
User1	10	4	2	2	8	5	5.17
User2	10	8	4	4	8	4	6.33
User3	10	10	6	8	10	10	9.00
User4	-	-	-	-	-	-	
User5	10	-	-	-	-	-	10.00
User6	8	8	2	2	3	5	4.67
User7	5	6	3	1	5	6	4.33
User8	8	8	6	10	10	10	8.67
User9	-	-	-	-	-	-	
User10	-	-	-	-	-	-	
Average	8.71	7.33	3.83	4.50	7.33	6.67	

Table E.32: Group 32 table of similarity ratings

E.2 Group Rating Graphs

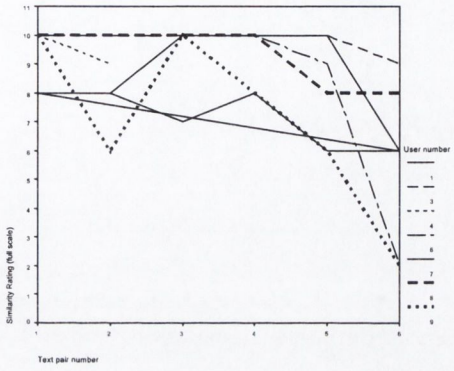


Figure E.1: Similarity Ratings for Group 1

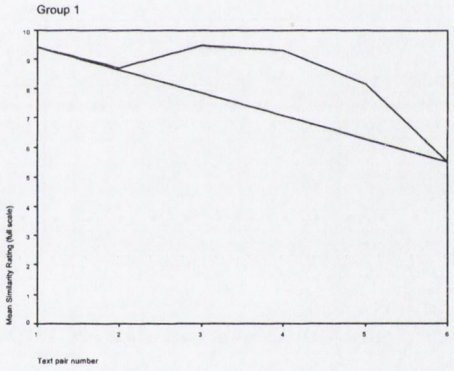


Figure E.2: Mean Similarity Ratings for Group 1

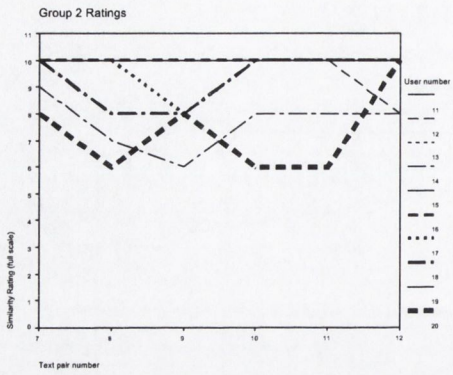


Figure E.3: Similarity Ratings for Group 2

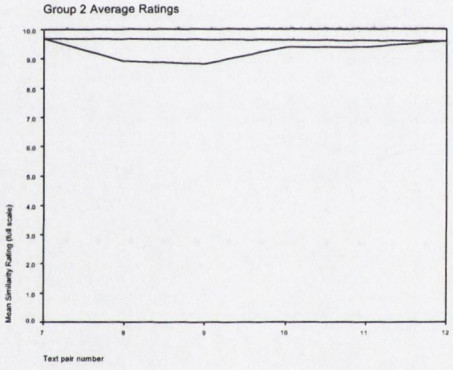


Figure E.4: Mean Similarity Ratings for Group 2

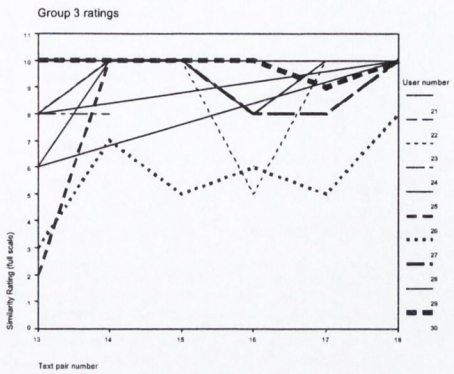


Figure E.5: Similarity Ratings for Group 3

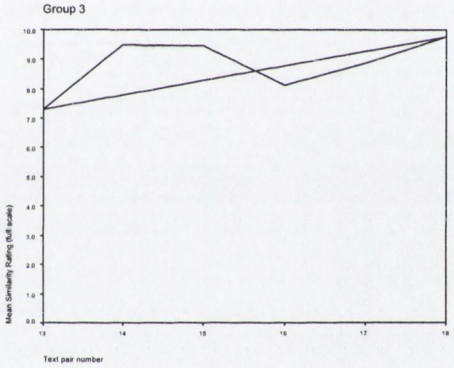


Figure E.6: Mean Similarity Ratings for Group 3

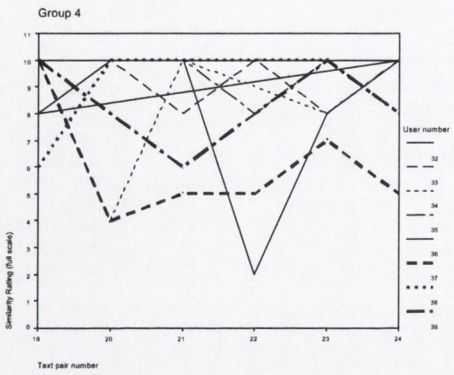


Figure E.7: Similarity Ratings for Group 4

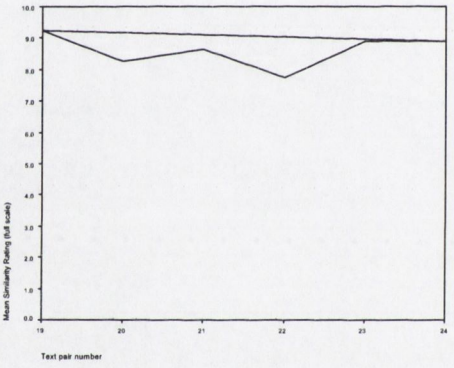


Figure E.8: Mean Similarity Ratings for Group 4

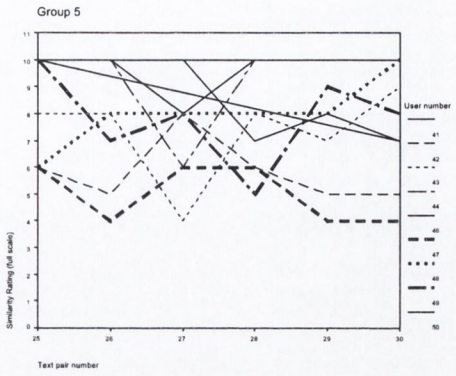


Figure E.9: Similarity Ratings for Group 5

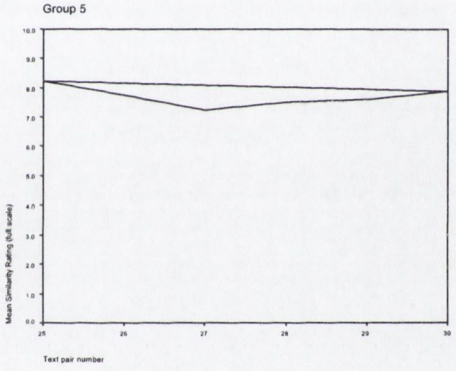


Figure E.10: Mean Similarity Ratings for Group 5

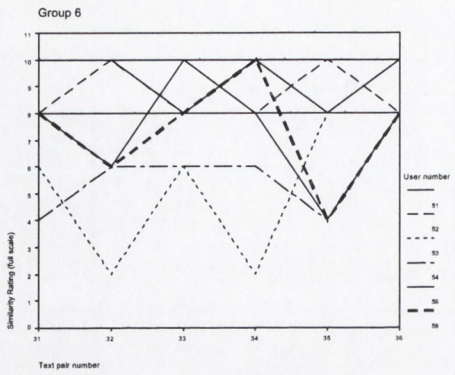


Figure E.11: Similarity Ratings for Group 6

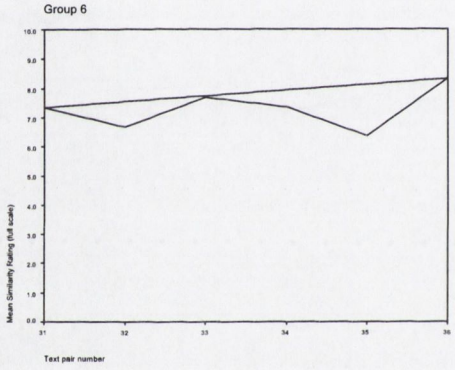


Figure E.12: Mean Similarity Ratings for Group 6

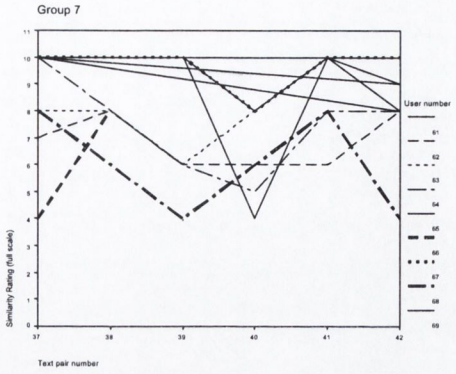


Figure E.13: Similarity Ratings for Group 7

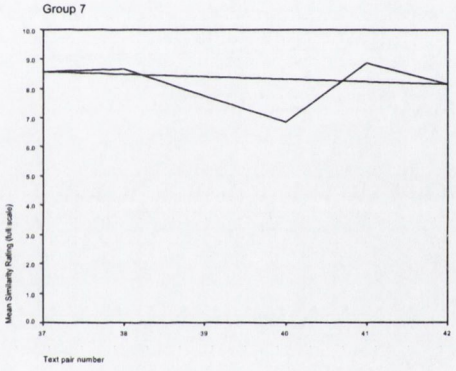


Figure E.14: Mean Similarity Ratings for Group 7

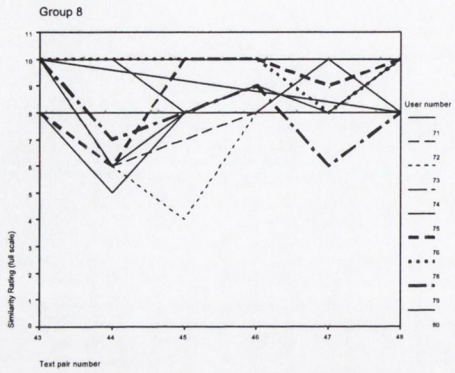


Figure E.15: Similarity Ratings for Group 8

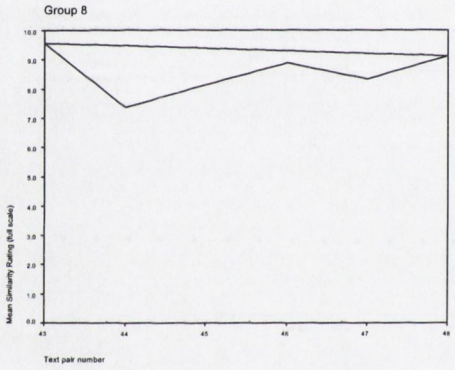


Figure E.16: Mean Similarity Ratings for Group 8

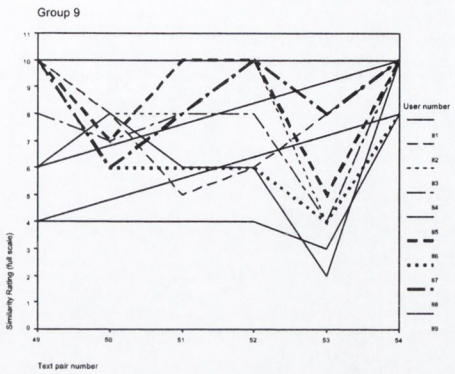


Figure E.17: Similarity Ratings for Group 9

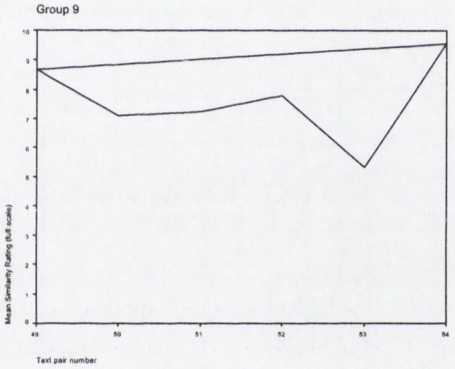


Figure E.18: Mean Similarity Ratings for Group 9

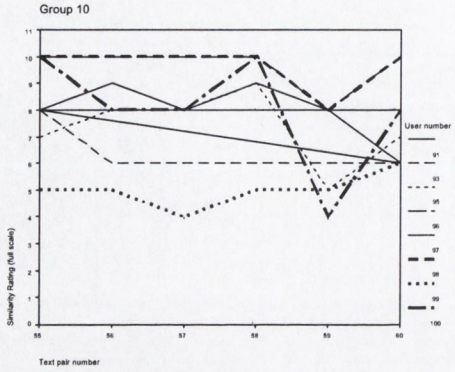


Figure E.19: Similarity Ratings for Group 10

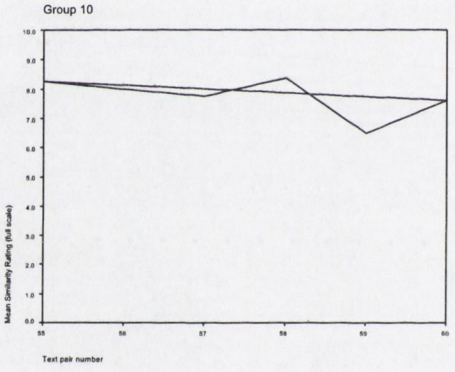


Figure E.20: Mean Similarity Ratings for Group 10

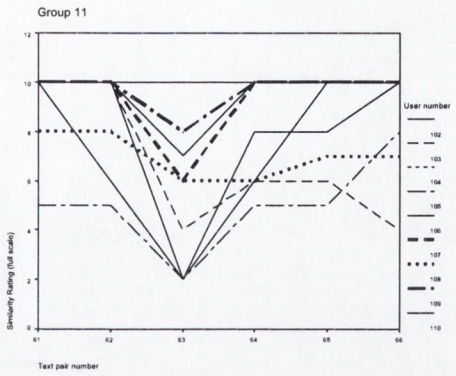


Figure E.21: Similarity Ratings for Group 11

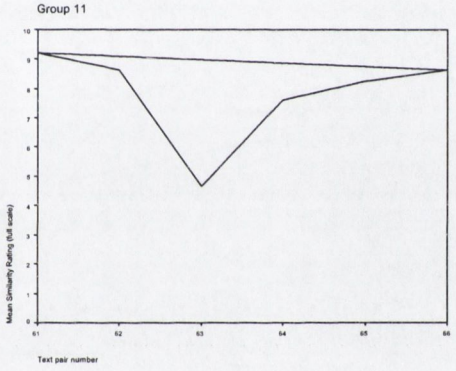


Figure E.22: Mean Similarity Ratings for Group 11

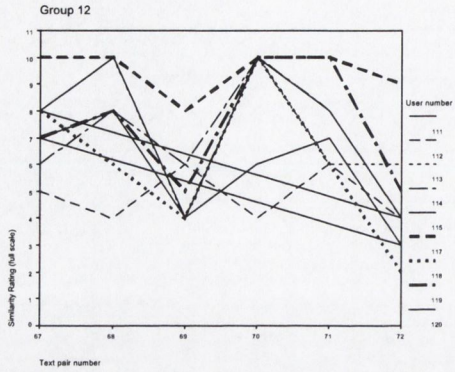


Figure E.23: Similarity Ratings for Group 12

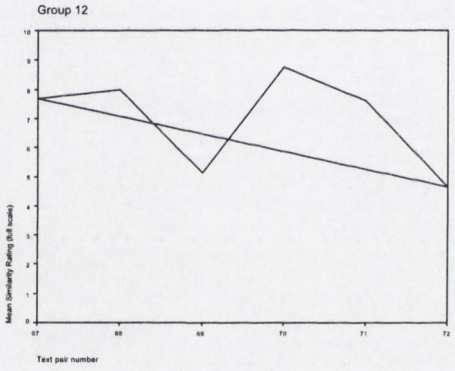


Figure E.24: Mean Similarity Ratings for Group 12

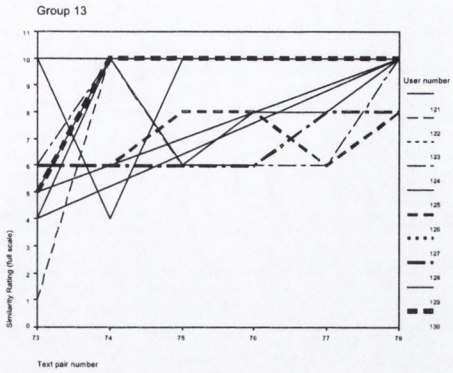


Figure E.25: Similarity Ratings for Group 13

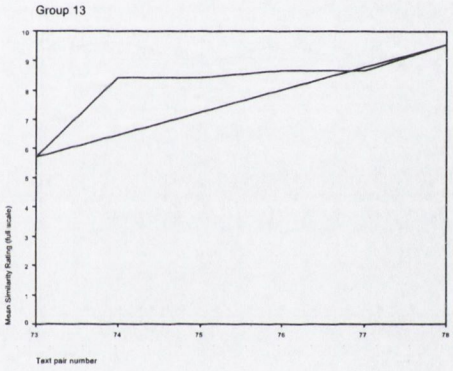


Figure E.26: Mean Similarity Ratings for Group 13

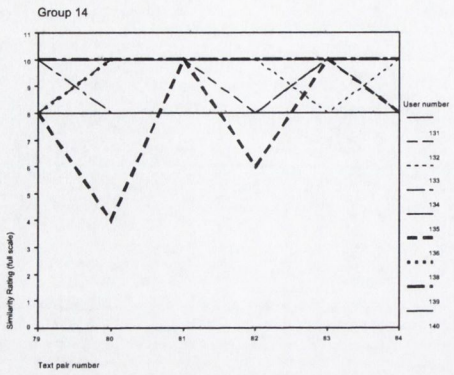


Figure E.27: Similarity Ratings for Group 14

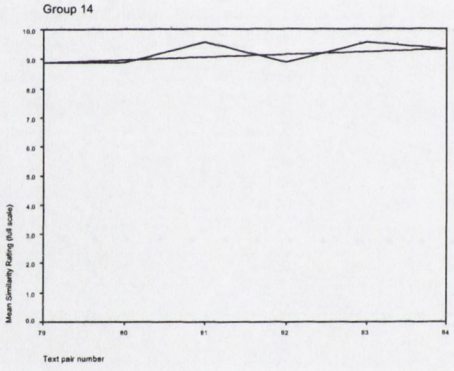


Figure E.28: Mean Similarity Ratings for Group 14

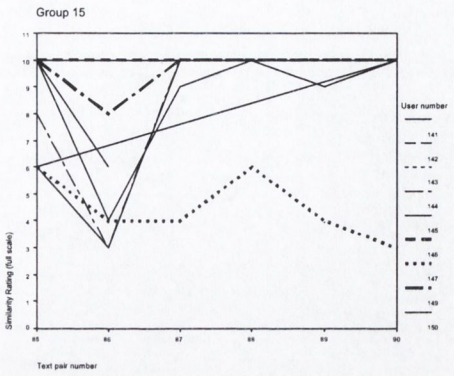


Figure E.29: Similarity Ratings for Group 15

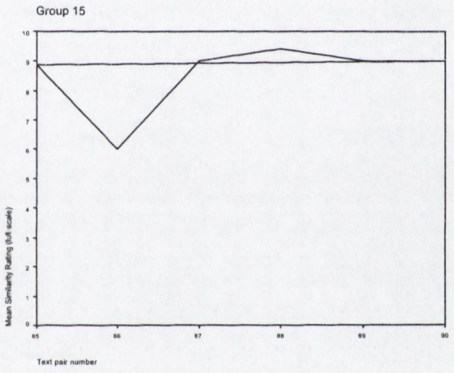


Figure E.30: Mean Similarity Ratings for Group 15

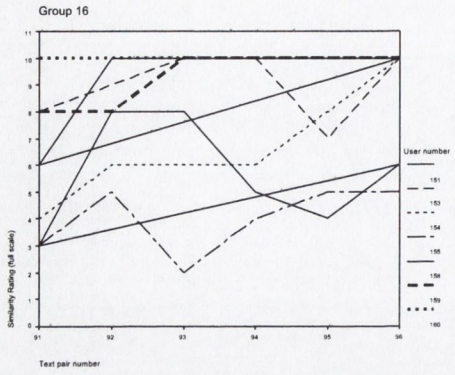


Figure E.31: Similarity Ratings for Group 16

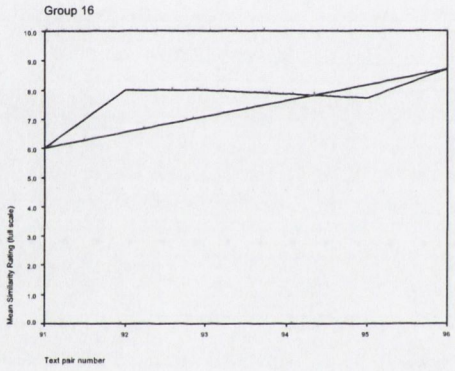


Figure E.32: Mean Similarity Ratings for Group 16

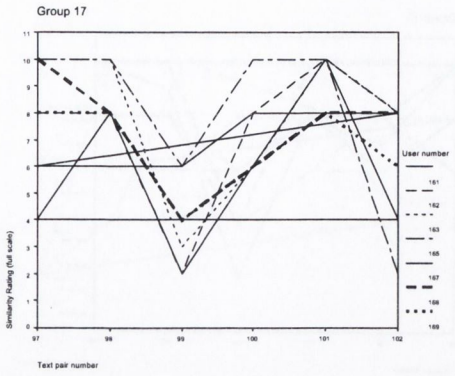


Figure E.33: Similarity Ratings for Group 17

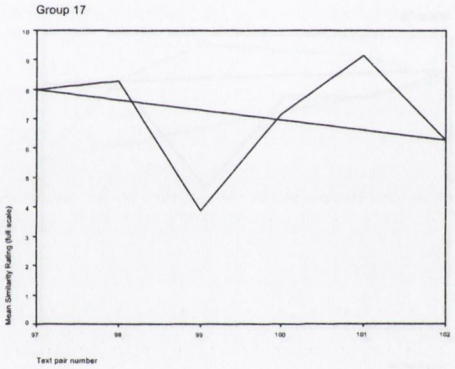


Figure E.34: Mean Similarity Ratings for Group 17

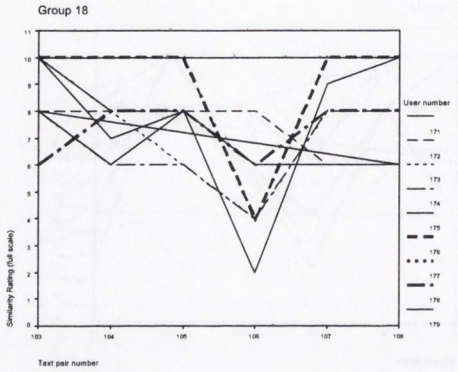


Figure E.35: Similarity Ratings for Group 17

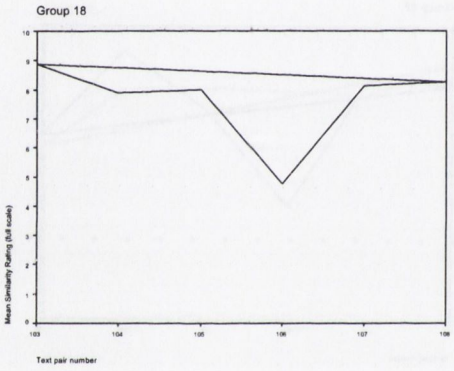


Figure E.36: Mean Similarity Ratings for Group 18

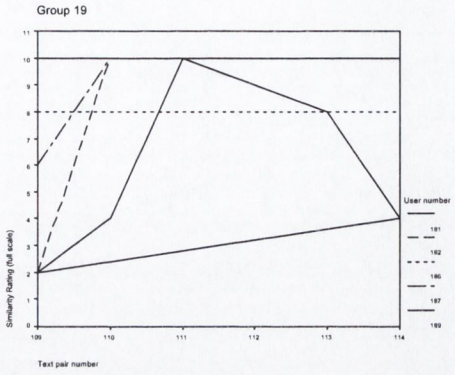


Figure E.37: Similarity Ratings for Group 19

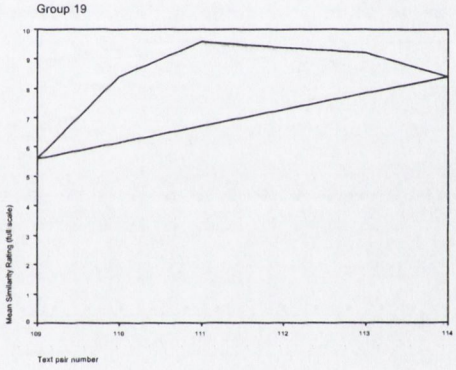


Figure E.38: Mean Similarity Ratings for Group 19

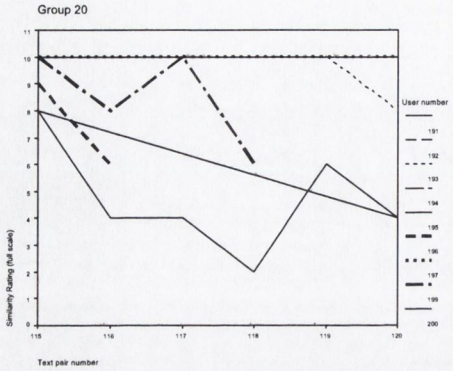


Figure E.39: Similarity Ratings for Group 20

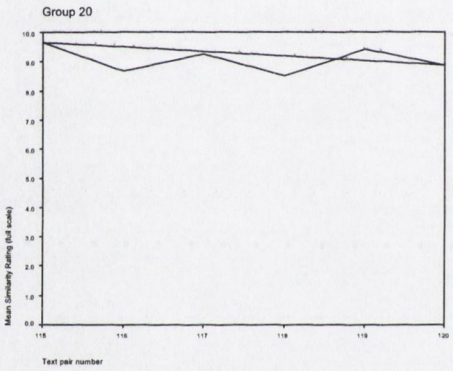


Figure E.40: Mean Similarity Ratings for Group 20

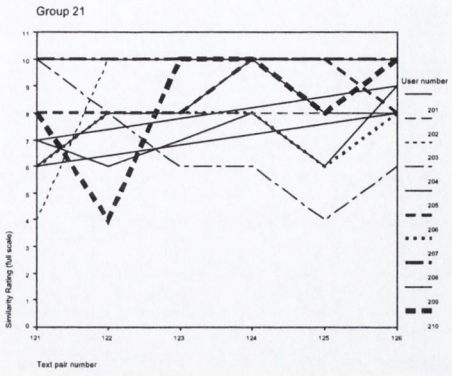


Figure E.41: Similarity Ratings for Group 21

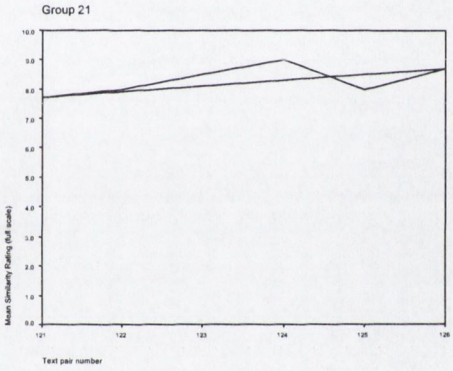


Figure E.42: Mean Similarity Ratings for Group 21

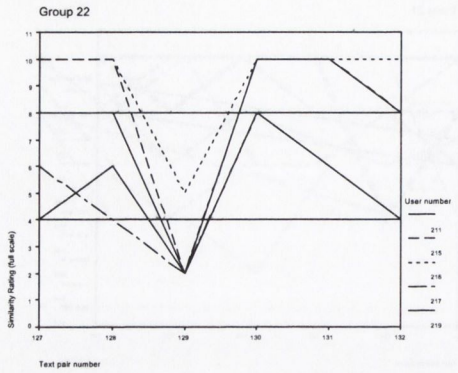


Figure E.43: Similarity Ratings for Group 22



Figure E.44: Mean Similarity Ratings for Group 22

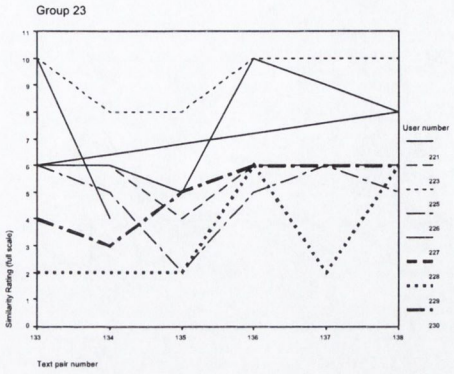


Figure E.45: Similarity Ratings for Group 23

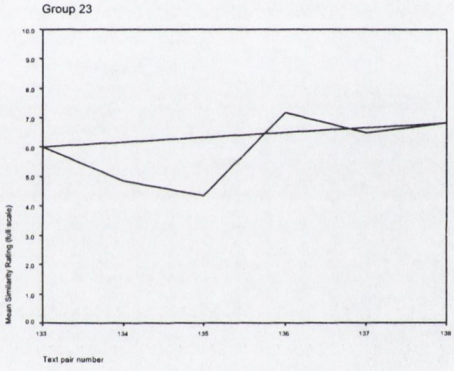


Figure E.46: Mean Similarity Ratings for Group 23

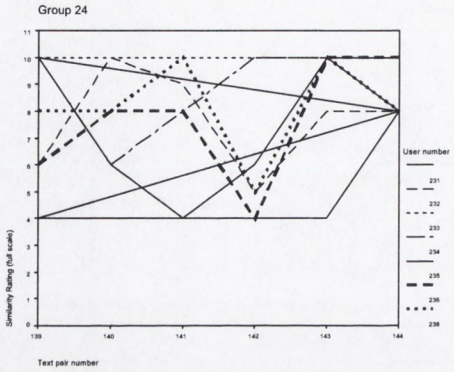


Figure E.47: Similarity Ratings for Group 24

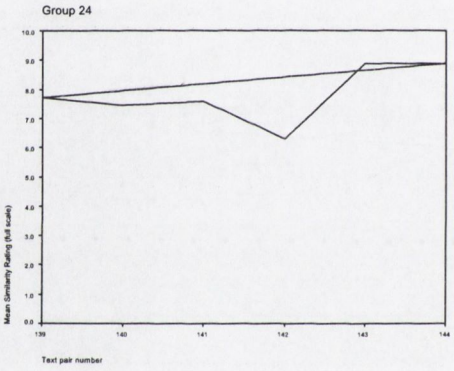


Figure E.48: Mean Similarity Ratings for Group 24

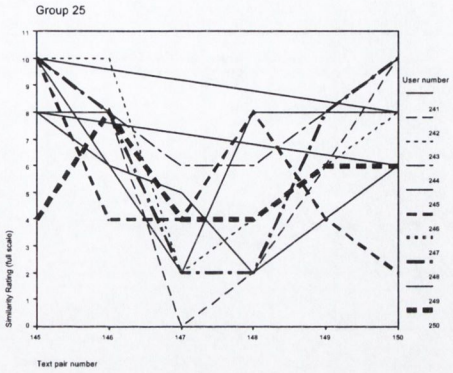


Figure E.49: Similarity Ratings for Group 25

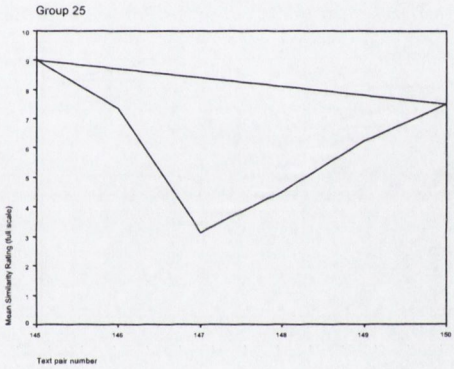


Figure E.50: Mean Similarity Ratings for Group 25

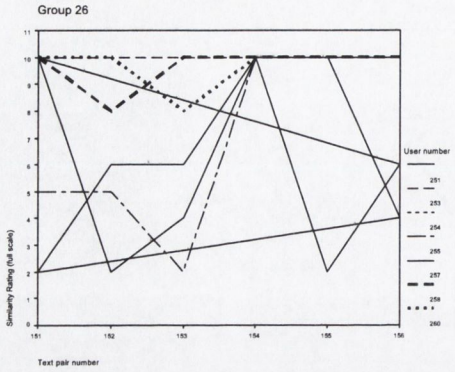


Figure E.51: Similarity Ratings for Group 26

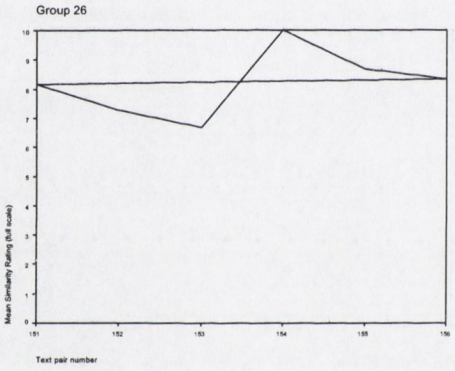


Figure E.52: Mean Similarity Ratings for Group 26

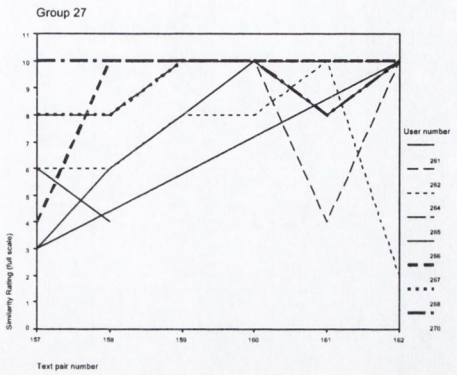


Figure E.53: Similarity Ratings for Group 27

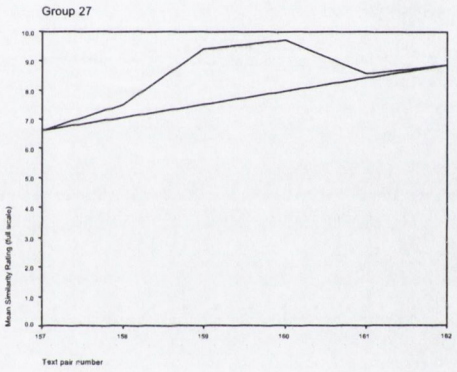


Figure E.54: Mean Similarity Ratings for Group 27

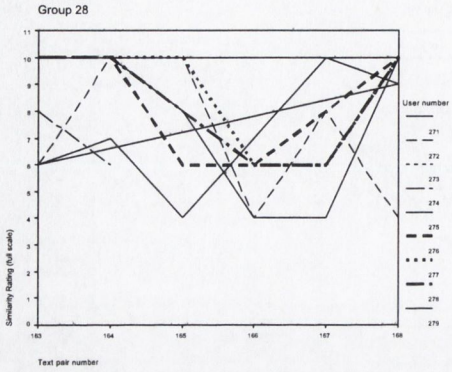


Figure E.55: Similarity Ratings for Group 28

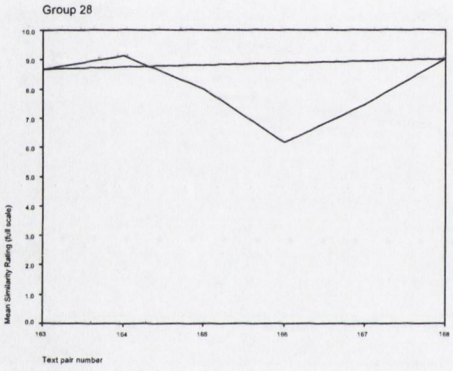


Figure E.56: Mean Similarity Ratings for Group 28

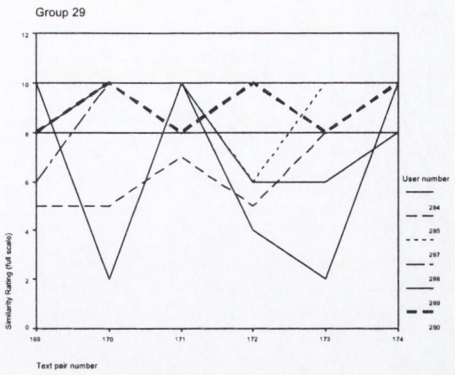


Figure E.57: Similarity Ratings for Group 29

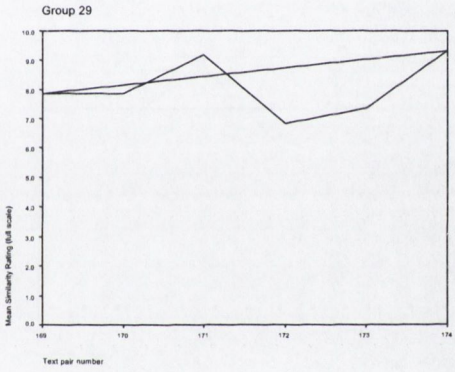


Figure E.58: Mean Similarity Ratings for Group 29

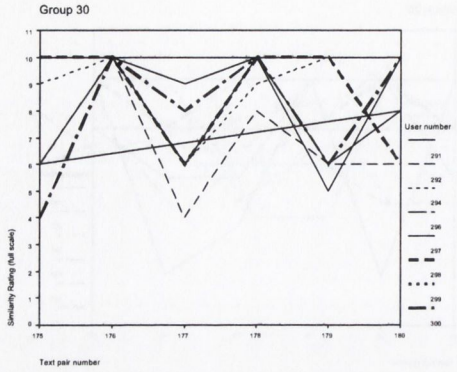


Figure E.59: Similarity Ratings for Group 30

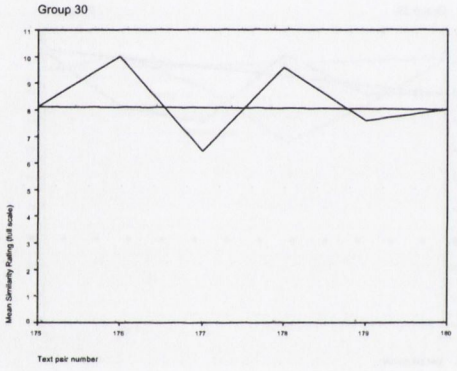


Figure E.60: Mean Similarity Ratings for Group 30

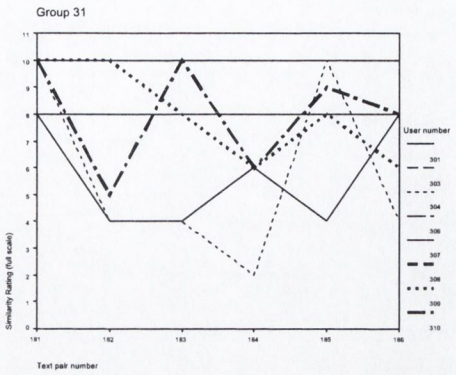


Figure E.61: Similarity Ratings for Group 31

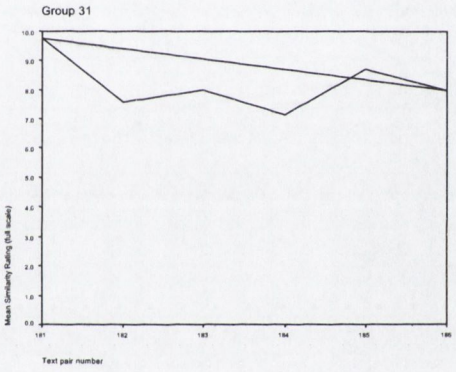


Figure E.62: Mean Similarity Ratings for Group 31

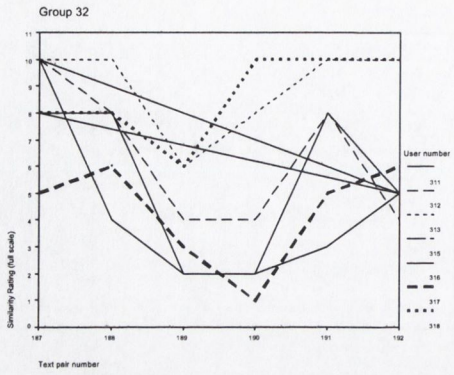


Figure E.63: Similarity Ratings for Group 32

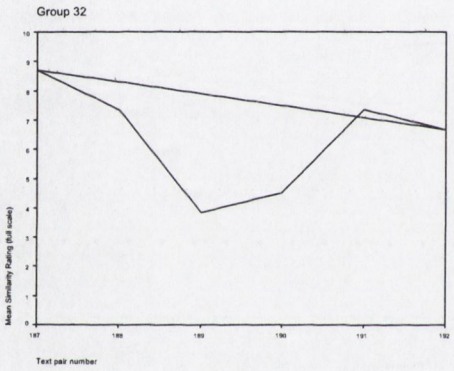


Figure E.64: Mean Similarity Ratings for Group 32

E.3 Siegel and Castellan Kappa Statistic Table

Story pair	0	1	2	3	4	5	S_i
1	0	0	0	0	2	5	0.5238
2	0	0	0	1	3	3	0.2857
3	0	0	0	1	0	5	0.6667
4	0	0	0	0	2	4	0.4667
5	0	0	0	2	2	2	0.2000
6	0	2	0	2	2	0	0.2000
7	0	0	0	0	2	8	0.6444
8	0	0	0	2	2	6	0.3778
9	0	0	0	1	4	5	0.3556
10	0	0	0	1	1	8	0.6222
11	0	0	0	1	1	8	0.6222
12	0	0	0	0	2	8	0.6444
13	0	2	0	1	4	3	0.2222
14	0	0	0	1	1	8	0.6222
15	0	0	1	0	0	8	0.7778
16	0	0	1	1	4	3	0.2500
17	0	0	1	0	3	5	0.3611
18	0	0	0	0	1	8	0.7778
19	0	0	0	1	1	6	0.5357
20	0	0	2	0	1	5	0.3929
21	0	0	1	1	1	5	0.3571
22	0	1	1	0	3	3	0.2143
23	0	0	0	1	3	4	0.3214
24	0	0	1	0	2	5	0.3929
25	0	0	0	3	2	4	0.2778
26	0	0	2	1	2	3	0.1786
27	0	0	1	2	4	1	0.2500
28	0	0	1	3	2	2	0.1786
29	0	0	2	1	3	2	0.1786
30	0	0	2	1	2	3	0.1786
31	0	0	1	1	3	1	0.2000
32	0	1	0	3	0	2	0.2667
33	0	0	0	2	3	1	0.2667
34	0	1	0	1	2	2	0.1333
35	0	0	3	0	2	1	0.2667
36	0	0	0	0	5	1	0.6667
37	0	0	1	1	2	5	0.3056
38	0	0	0	1	4	4	0.3333
39	0	0	1	3	0	4	0.3214
40	0	0	2	2	3	1	0.1786
41	0	0	0	1	2	4	0.3333
42	0	0	1	0	4	2	0.3333
43	0	0	0	0	2	7	0.6111
44	0	0	1	5	0	3	0.3611

45	0	0	1	1	4	3	0.2500
46	0	0	0	0	6	3	0.5000
47	0	0	0	1	6	2	0.4444
48	0	0	0	0	4	5	0.4444
49	0	0	1	1	1	6	0.4167
50	0	0	1	4	3	1	0.2500
51	0	0	2	2	3	2	0.1667
52	0	0	1	3	1	4	0.2500
53	0	2	4	0	2	1	0.2222
54	0	0	0	0	2	7	0.6111
55	0	0	1	1	3	3	0.2143
56	0	0	1	1	4	2	0.2500
57	0	0	1	1	4	2	0.2500
58	0	0	1	1	3	3	0.2143
59	0	0	3	1	4	0	0.3214
60	0	0	0	4	2	2	0.2857
61	0	0	1	0	1	7	0.5833
62	0	0	1	1	1	5	0.3571
63	0	3	1	3	1	0	0.2143
64	0	0	1	3	1	3	0.2143
65	0	0	1	2	1	4	0.2500
66	0	0	1	1	1	5	0.3571
67	0	0	1	3	3	2	0.1944
68	0	0	1	1	3	3	0.2143
69	0	0	5	2	1	0	0.3929
70	0	0	1	1	0	6	0.5357
71	0	0	0	4	2	2	0.2857
72	0	2	4	1	1	0	0.2500
73	1	0	4	3	0	2	0.2222
74	0	0	1	2	0	6	0.4444
75	0	0	0	3	1	5	0.3611
76	0	0	0	2	2	5	0.3333
77	0	0	0	2	2	5	0.3333
78	0	0	0	0	2	7	0.6111
79	0	0	0	0	5	4	0.4444
80	0	0	1	0	2	6	0.4444
81	0	0	0	0	2	7	0.6111
82	0	0	0	1	3	5	0.3611
83	0	0	0	0	2	7	0.6111
84	0	0	0	0	3	6	0.5000
85	0	0	0	2	1	6	0.4444
86	0	2	2	1	1	2	0.1071
87	0	0	1	0	1	5	0.4762
88	0	0	0	1	0	6	0.7143
89	0	0	1	0	1	5	0.4762
90	0	1	0	0	0	6	0.7143
91	0	2	1	1	2	1	0.0952
92	0	0	1	1	3	2	0.1905
93	0	1	0	1	1	4	0.2857

94	0	0	2	1	0	4	0.3333
95	0	0	2	1	1	3	0.1905
96	0	0	1	1	0	5	0.4762
97	0	0	1	1	2	3	0.1905
98	0	0	0	1	4	2	0.3333
99	0	3	2	2	0	0	0.2381
100	0	0	0	4	2	1	0.3333
101	0	0	0	0	3	4	0.4286
102	0	1	1	1	4	0	0.2857
103	0	0	0	1	3	5	0.3611
104	0	0	0	3	4	2	0.2778
105	0	0	0	2	4	2	0.2857
106	0	1	4	2	1	0	0.2500
107	0	0	0	2	4	2	0.2857
108	0	0	0	2	3	3	0.2500
109	0	2	0	1	1	1	0.1000
110	0	0	1	0	1	3	0.3000
111	0	0	0	0	1	4	0.6000
112	0	0	0	0	2	3	0.4000
113	0	0	0	0	2	3	0.4000
114	0	0	1	0	1	3	0.3000
115	0	0	0	0	2	7	0.6111
116	0	0	1	1	1	6	0.4167
117	0	0	1	0	0	7	0.7500
118	0	1	0	1	0	6	0.5357
119	0	0	0	1	0	6	0.7143
120	0	0	1	0	1	5	0.4762
121	0	0	1	3	3	3	0.2000
122	0	0	1	1	5	3	0.2889
123	0	0	0	2	4	4	0.2889
124	0	0	0	1	3	6	0.4000
125	0	0	1	2	3	4	0.2222
126	0	0	0	1	5	4	0.3556
127	0	0	1	1	1	2	0.1000
128	0	0	1	1	1	2	0.1000
129	0	4	1	0	0	0	0.6000
130	0	0	0	0	2	3	0.4000
131	0	0	0	1	1	3	0.3000
132	0	0	1	0	3	1	0.3000
133	0	1	2	3	0	2	0.1786
134	0	2	2	2	1	0	0.1429
135	0	2	3	0	1	0	0.2667
136	0	0	1	3	0	2	0.2667
137	0	1	0	3	1	1	0.2000
138	0	0	1	3	1	1	0.2000
139	0	0	1	2	1	3	0.1905
140	0	0	1	2	2	2	0.1429
141	0	0	2	0	3	2	0.2381
142	0	0	4	1	0	2	0.3333

143	0	0	1	0	1	5	0.4762
144	0	0	0	0	4	3	0.4286
145	0	0	1	0	2	7	0.4889
146	0	0	1	2	5	1	0.3056
147	1	3	3	1	0	0	0.2143
148	0	3	2	1	2	0	0.1786
149	0	0	2	3	3	0	0.2500
150	0	1	0	2	2	3	0.1786
151	0	1	1	0	0	5	0.4762
152	0	1	1	1	1	3	0.1429
153	0	1	1	1	1	2	0.0667
154	0	0	0	0	0	6	1.0000
155	0	1	0	0	0	5	0.6667
156	0	0	1	1	0	4	0.4000
157	0	1	1	2	3	1	0.1429
158	0	0	1	2	3	2	0.1786
159	0	0	0	0	2	5	0.5238
160	0	0	0	0	1	6	0.7143
161	0	0	1	0	2	4	0.3333
162	0	1	0	0	0	6	0.7143
163	0	0	0	2	2	5	0.3333
164	0	0	0	2	0	6	0.5714
165	0	0	1	1	2	3	0.1905
166	0	0	2	4	0	1	0.3333
167	0	0	1	2	2	2	0.1429
168	0	0	1	0	1	5	0.4762
169	0	0	1	1	2	2	0.1333
170	0	1	1	0	0	4	0.4000
171	0	0	0	1	1	4	0.4000
172	0	0	2	2	0	2	0.2000
173	0	1	0	1	2	2	0.1333
174	0	0	0	0	2	4	0.4667
175	0	0	1	2	1	4	0.2500
176	0	0	0	0	0	7	1.0000
177	0	0	1	4	2	0	0.3333
178	0	0	0	0	2	5	0.5238
179	0	0	1	3	0	3	0.2857
180	0	0	0	3	1	3	0.2857
181	0	0	0	0	1	7	0.7500
182	0	0	3	0	0	4	0.4286
183	0	0	2	0	1	4	0.3333
184	0	1	0	3	0	3	0.2857
185	0	0	1	0	2	4	0.3333
186	0	0	1	1	2	3	0.1905
187	0	0	1	0	2	4	0.3333
188	0	0	1	1	3	1	0.2000
189	0	3	1	2	0	0	0.2667
190	1	2	1	0	1	1	0.0667
191	0	1	1	0	2	2	0.1333

192	0	0	3	1	0	2	0.2667
$Total_j$	3	60	159	223	350	668	
p_j	0.0021	0.0410	0.1087	0.1524	0.2392	0.4566	

E.4 Euclidean Distance from Matlab Function *pdist*

$Y = pdist(X)$ computes the Euclidean distance between pairs of objects in m-by-n matrix X, which is treated as m vectors of size n. For a data set made up of m objects, there are $\frac{m(m-1)}{2}$ pairs.

The output, Y, is a vector of length $\frac{m(m-1)}{2}$, containing the distance information. The distances are arranged in the order (1,2), (1,3), ..., (1,m), (2,3), ..., (2,m), ..., ..., (m-1,m). Y is also commonly known as a similarity matrix or dissimilarity matrix.

Given an m-by-n data matrix X, which is treated as m (1-by-n) row vectors x1, x2, ..., xm, the various distances between the vector xr and xs are defined as follows:

Euclidean distance

$$d_{rs}^2 = (x_r - x_s)(x_r - x_s)' \tag{E.1}$$

E.4.1 Tests with artificial data: all different

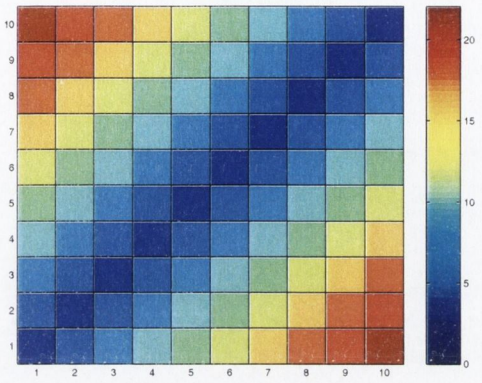


Figure E.65: Euclidean Distance Matrix: all different

User	Similarity Rating (from zero to ten)					
10	10	10	10	10	10	10
9	9	9	9	9	9	9
8	8	8	8	8	8	8
7	7	7	7	7	7	7
6	6	6	6	6	6	6
5	5	5	5	5	5	5
4	4	4	4	4	4	4
3	3	3	3	3	3	3
2	2	2	2	2	2	2
1	1	1	1	1	1	1

User	Euclidean Distance									
10	22	20	17	15	12	10	7	5	2	0
9	20	17	15	12	10	7	5	2	0	2
8	17	15	12	10	7	5	2	0	2	5
7	15	12	10	7	5	2	0	2	5	7
6	12	10	7	5	2	0	2	5	7	10
5	10	7	5	2	0	2	5	7	10	12
4	7	5	2	0	2	5	7	10	12	15
3	5	2	0	2	5	7	10	12	15	17
2	2	0	2	5	7	10	12	15	17	20
1	0	2	5	7	10	12	15	17	20	22

Euclidean distance: all 1s

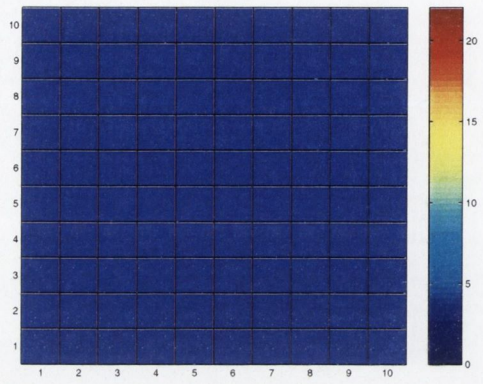


Figure E.66: Euclidean Distance Matrix: all 1s

User	Similarity Rating (from zero to ten)					
10	1	1	1	1	1	1
9	1	1	1	1	1	1
8	1	1	1	1	1	1
7	1	1	1	1	1	1
6	1	1	1	1	1	1
5	1	1	1	1	1	1
4	1	1	1	1	1	1
3	1	1	1	1	1	1
2	1	1	1	1	1	1
1	1	1	1	1	1	1

User	Euclidean Distance									
10	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0

Euclidean distance: all 10s

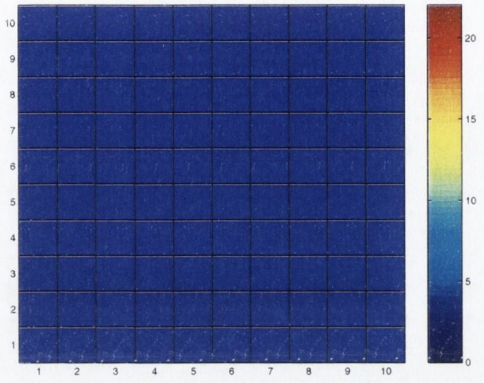


Figure E.67: Euclidean Distance Matrix: all 10s

User	Similarity Rating (from zero to ten)					
10	10	10	10	10	10	10
9	10	10	10	10	10	10
8	10	10	10	10	10	10
7	10	10	10	10	10	10
6	10	10	10	10	10	10
5	10	10	10	10	10	10
4	10	10	10	10	10	10
3	10	10	10	10	10	10
2	10	10	10	10	10	10
1	10	10	10	10	10	10

User	Euclidean Distance									
10	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0

E.4.2 Tests with user data

Group 1

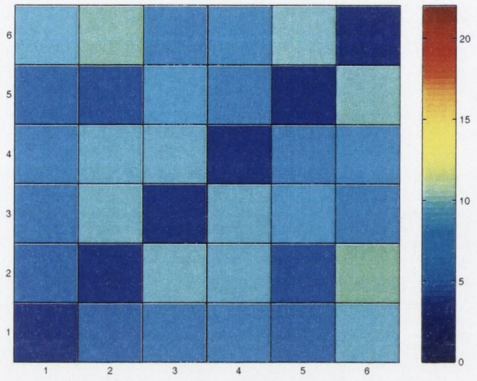


Figure E.68: Group 1 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
6	10	6	10	8	6	2
5	10	10	10	10	8	8
4	8	8	7	8	6	6
3	10	10	10	10	9	2
2	10	10	10	10	10	9
1	8	8	10	10	10	6

User	Euclidean Distances					
6	6.6	9.2	5.4	5.7	7.7	0.0
5	4.0	2.2	6.1	5.4	0.0	7.7
4	5.4	6.8	6.8	0.0	5.4	5.7
3	5.0	7.1	0.0	6.8	6.1	5.4
2	4.1	0.0	7.1	6.8	2.2	9.2
1	0.0	4.1	5.0	5.4	4.0	6.6

Group 2

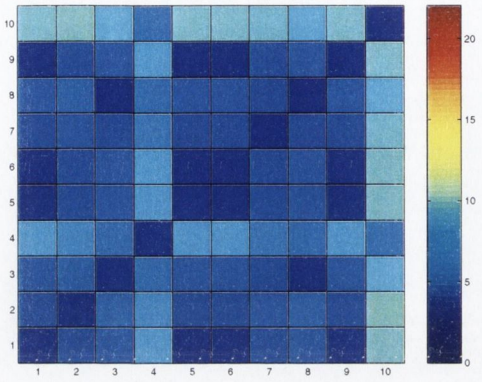


Figure E.69: Group 2 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
10	8	6	8	6	6	10
9	10	10	10	10	10	10
8	10	8	8	10	10	10
7	10	10	8	10	10	10
6	10	10	10	10	10	10
5	10	10	10	10	10	10
4	9	7	6	8	8	8
3	10	8	8	10	10	10
2	10	10	10	10	10	8
1	10	10	10	10	10	10

User	Euclidean Distances									
10	7.5	7.7	6.3	4.2	7.5	7.5	7.2	6.3	7.5	0.0
9	0.0	2.0	2.8	6.2	0.0	0.0	2.0	2.8	0.0	7.5
8	2.8	3.5	0.0	4.2	2.8	2.8	2.0	0.0	2.8	6.3
7	2.0	2.8	2.0	5.1	2.0	2.0	0.0	2.0	2.0	7.2
6	0.0	2.0	2.8	6.2	0.0	0.0	2.0	2.8	0.0	7.5
5	0.0	2.0	2.8	6.2	0.0	0.0	2.0	2.8	0.0	7.5
4	6.2	5.8	4.2	0.0	6.2	6.2	5.1	4.2	6.2	4.2
3	2.8	3.5	0.0	4.2	2.8	2.8	2.0	0.0	2.8	6.3
2	2.0	0.0	3.5	5.8	2.0	2.0	2.8	3.5	2.0	7.7
1	0.0	2.0	2.8	6.2	0.0	0.0	2.0	2.8	0.0	7.5

Group 3

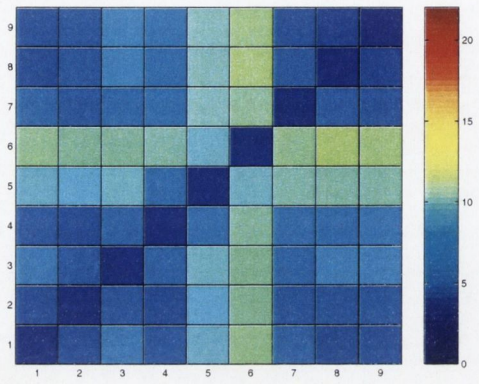


Figure E.70: Group 3 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
9	10	10	10	10	9	10
8	10	10	10	10	10	10
7	10	10	10	8	8	10
6	3	7	5	6	5	8
5	2	10	10	8	8	10
4	6	10	10	8	10	10
3	8	10	10	5	10	10
2	8	10	10	8	10	10
1	8	10	10	10	10	10

User	Euclidean Distances								
9	2.2	3.0	5.5	4.6	8.3	10.9	2.2	1.0	0.0
8	2.0	2.8	5.4	4.5	8.5	11.3	2.8	0.0	1.0
7	3.5	2.8	4.1	4.5	8.0	10.0	0.0	2.8	2.2
6	10.2	9.6	9.4	8.7	7.2	0.0	10.0	11.3	10.9
5	6.6	6.3	7.0	4.5	0.0	7.2	8.0	8.5	8.3
4	2.8	2.0	3.6	0.0	4.5	8.7	4.5	4.5	4.6
3	5.0	3.0	0.0	3.6	7.0	9.4	4.1	5.4	5.5
2	2.0	0.0	3.0	2.0	6.3	9.6	2.8	2.8	3.0
1	0.0	2.0	5.0	2.8	6.6	10.2	3.5	2.0	2.2

Group 4

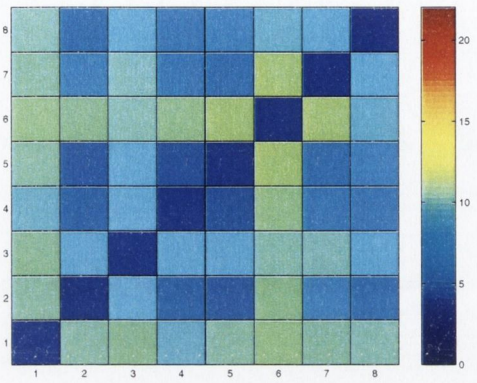


Figure E.71: Group 4 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
8	10	8	6	8	10	8
7	6	10	10	10	10	8
6	10	4	5	5	7	5
5	10	10	10	10	10	10
4	10	10	10	8	10	10
3	10	4	10	9	8	10
2	10	10	8	10	8	10
1	8	10	10	2	8	10

User	Euclidean Distances							
8	8.2	4.5	6.4	4.9	5.3	6.6	6.3	0.0
7	8.7	5.3	7.8	4.9	4.5	11.0	0.0	6.3
6	10.0	9.8	8.2	10.2	11.0	0.0	11.0	6.6
5	8.5	2.8	6.4	2.0	0.0	11.0	4.5	5.3
4	6.6	3.5	6.4	0.0	2.0	10.2	4.9	4.9
3	9.4	6.4	0.0	6.4	6.4	8.2	7.8	6.4
2	8.5	0.0	6.4	3.5	2.8	9.8	5.3	4.5
1	0.0	8.5	9.4	6.6	8.5	10.0	8.7	8.2

Group 5

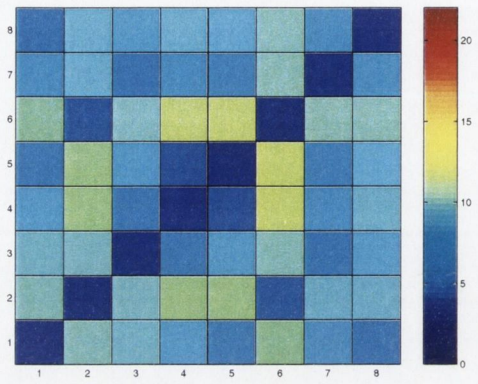


Figure E.72: Group 5 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
8	10	7	8	5	9	8
7	6	8	8	8	8	10
6	6	4	6	6	4	4
5	10	10	8	10	10	10
4	10	10	6	10	10	10
3	8	8	4	8	7	9
2	6	5	8	6	5	5
1	10	10	10	7	8	7

User	Euclidean Distances							
8	4.4	6.8	5.9	6.6	6.2	8.4	5.6	0.0
7	5.8	6.9	4.7	5.7	5.3	8.7	0.0	5.6
6	9.7	2.6	7.9	11.8	12.0	0.0	8.7	8.4
5	5.1	10.3	6.2	2.0	0.0	12.0	5.3	6.2
4	6.2	10.5	5.1	0.0	2.0	11.8	5.7	6.6
3	7.1	7.3	0.0	5.1	6.2	7.9	4.7	5.9
2	7.7	0.0	7.3	10.5	10.3	2.6	6.9	6.8
1	0.0	7.7	7.1	6.2	5.1	9.7	5.8	4.4

Group 6

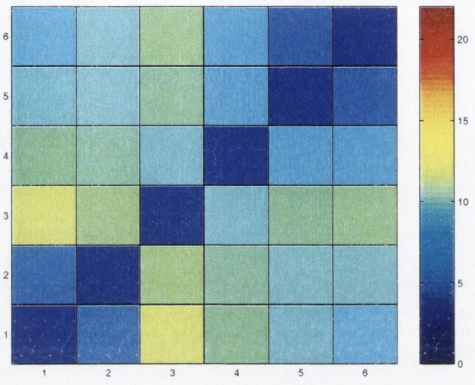


Figure E.73: Group 6 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
6	8	6	8	10	4	8
5	8	6	10	8	4	8
4	4	6	6	6	4	8
3	6	2	6	2	8	8
2	8	10	8	8	10	8
1	10	10	8	10	8	10

User	Euclidean Distances					
6	6.3	7.5	10.2	6.0	2.8	0.0
5	6.9	7.5	9.4	6.0	0.0	2.8
4	9.6	8.7	7.2	0.0	6.0	6.0
3	12.3	10.6	0.0	7.2	9.4	10.2
2	4.0	0.0	10.6	8.7	7.5	7.5
1	0.0	4.0	12.3	9.6	6.9	6.3

Group 7

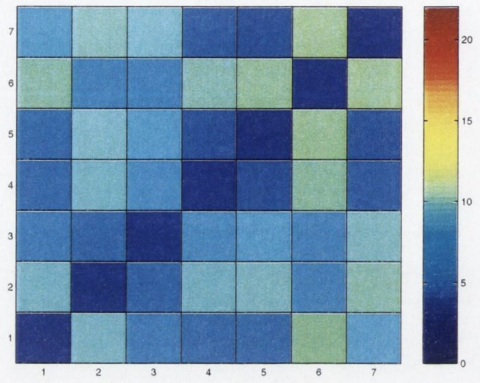


Figure E.74: Group 7 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
7	10	10	10	10	10	10
6	8	6	4	6	8	4
5	10	10	10	8	10	10
4	10	10	10	8	10	8
3	10	8	6	5	8	8
2	7	8	6	6	6	8
1	10	10	10	4	10	9

User	Euclidean Distances						
7	6.1	8.1	7.3	2.8	2.0	10.6	0.0
6	9.4	5.4	5.4	8.9	10.0	0.0	10.6
5	4.1	7.3	6.1	2.0	0.0	10.0	2.0
4	4.1	7.0	5.7	0.0	2.0	8.9	2.8
3	5.1	3.7	0.0	5.7	6.1	5.4	7.3
2	7.1	0.0	3.7	7.0	7.3	5.4	8.1
1	0.0	7.1	5.1	4.1	4.1	9.4	6.1

Group 8

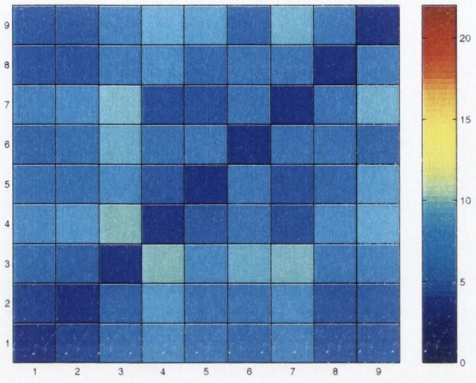


Figure E.75: Group 8 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
9	8	5	8	8	10	8
8	10	7	8	9	6	8
7	10	10	10	10	8	10
6	8	6	10	10	9	10
5	10	10	8	9	8	10
4	10	10	10	10	10	10
3	10	6	4	8	8	10
2	10	6	7	8	8	8
1	10	6	8	8	8	8

User	Euclidean Distances								
9	3.0	3.2	5.4	6.4	6.2	3.7	6.7	5.0	0.0
8	2.4	2.6	5.1	5.8	4.1	4.8	4.7	0.0	5.0
7	5.3	5.7	7.5	2.0	2.2	4.6	0.0	4.7	6.7
6	4.1	4.7	6.7	4.6	5.1	0.0	4.6	4.8	3.7
5	4.6	4.7	5.7	3.0	0.0	5.1	2.2	4.1	6.2
4	5.7	6.1	7.7	0.0	3.0	4.6	2.0	5.8	6.4
3	4.5	3.6	0.0	7.7	5.7	6.7	7.5	5.1	5.4
2	1.0	0.0	3.6	6.1	4.7	4.7	5.7	2.6	3.2
1	0.0	1.0	4.5	5.7	4.6	4.1	5.3	2.4	3.0

Group 9

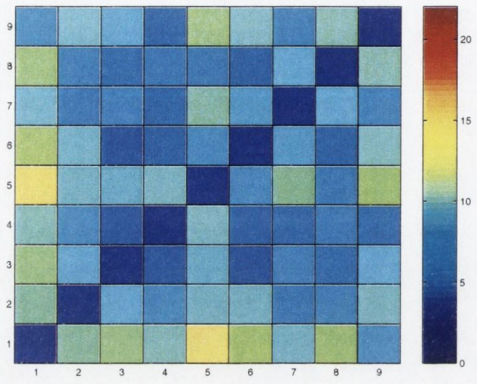


Figure E.76: Group 9 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
9	6	8	6	6	2	10
8	10	6	8	10	8	10
7	10	6	6	6	4	8
6	10	7	10	10	5	10
5	10	10	10	10	10	10
4	8	7	8	8	4	10
3	10	8	8	10	4	10
2	10	8	5	6	8	10
1	4	4	4	4	3	8

User	Euclidean Distances								
9	5.7	7.3	6.3	4.1	10.8	7.6	5.3	8.7	0.0
8	11.0	5.4	4.5	5.0	4.9	3.7	6.3	0.0	8.7
7	7.0	5.0	5.3	4.1	9.4	6.2	0.0	6.3	5.3
6	11.2	7.1	2.4	3.6	5.8	0.0	6.2	3.7	7.6
5	14.0	7.0	6.6	7.5	0.0	5.8	9.4	4.9	10.8
4	7.9	5.8	3.0	0.0	7.5	3.6	4.1	5.0	4.1
3	10.4	6.4	0.0	3.0	6.6	2.4	5.3	4.5	6.3
2	9.3	0.0	6.4	5.8	7.0	7.1	5.0	5.4	7.3
1	0.0	9.3	10.4	7.9	14.0	11.2	7.0	11.0	5.7

Group 10

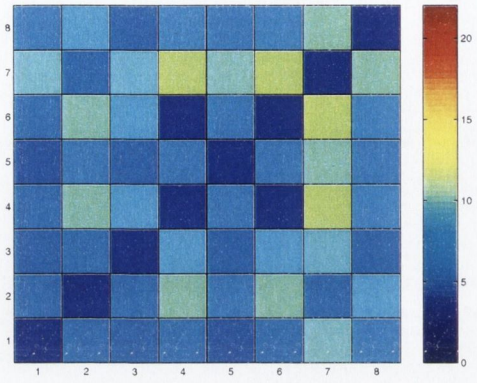


Figure E.77: Group 10 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
8	10	8	8	10	4	8
7	5	5	4	5	5	6
6	10	10	10	10	8	10
5	8	9	8	9	8	6
4	10	10	10	10	8	10
3	7	8	8	9	5	7
2	8	6	6	6	6	6
1	8	8	8	8	8	8

User	Euclidean Distances							
8	4.9	6.0	3.5	5.3	5.1	5.3	8.9	0.0
7	7.5	4.0	6.8	11.7	8.1	11.7	0.0	8.9
6	4.5	8.5	6.0	0.0	5.1	0.0	11.7	5.3
5	2.4	5.1	3.5	5.1	0.0	5.1	8.1	5.1
4	4.5	8.5	6.0	0.0	5.1	0.0	11.7	5.3
3	3.5	4.5	0.0	6.0	3.5	6.0	6.8	3.5
2	4.5	0.0	4.5	8.5	5.1	8.5	4.0	6.0
1	0.0	4.5	3.5	4.5	2.4	4.5	7.5	4.9

Group 11

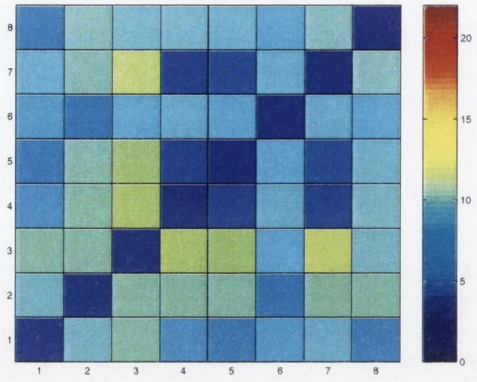


Figure E.78: Group 11 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
8	10	6	2	6	10	10
7	10	10	8	10	10	10
6	8	8	6	6	7	7
5	10	10	6	10	10	10
4	10	10	7	10	10	10
3	5	5	2	5	5	8
2	10	10	4	6	6	4
1	10	10	2	8	8	10

User	Euclidean Distances							
8	4.9	8.5	7.5	7.5	6.9	6.5	8.2	0.0
7	6.6	9.2	11.8	1.0	2.0	6.8	0.0	8.2
6	6.2	4.7	6.3	6.6	6.5	0.0	6.8	6.5
5	4.9	8.5	11.0	1.0	0.0	6.5	2.0	6.9
4	5.7	8.8	11.4	0.0	1.0	6.6	1.0	7.5
3	8.5	8.5	0.0	11.4	11.0	6.3	11.8	7.5
2	6.9	0.0	8.5	8.8	8.5	4.7	9.2	8.5
1	0.0	6.9	8.5	5.7	4.9	6.2	6.6	4.9

Group 12

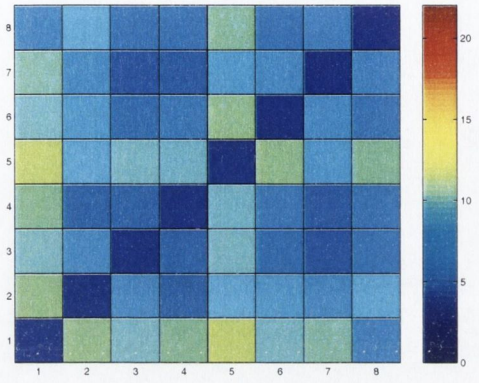


Figure E.79: Group 12 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
8	7	8	4	6	7	3
7	7	8	5	10	10	5
6	8	6	4	10	6	2
5	10	10	8	10	10	9
4	8	10	4	10	8	4
3	6	8	6	10	8	4
2	10	10	4	10	6	6
1	5	4	6	4	6	4

User	Euclidean Distances							
8	5.5	6.2	4.8	4.8	9.5	4.8	5.5	0.0
7	8.6	5.6	2.6	3.3	6.2	5.6	0.0	5.5
6	7.5	6.0	4.5	4.9	10.0	0.0	5.6	4.8
5	11.9	6.4	7.3	7.0	0.0	10.0	6.2	9.5
4	9.4	3.5	3.5	0.0	7.0	4.9	3.3	4.8
3	7.5	5.7	0.0	3.5	7.3	4.5	2.6	4.8
2	10.2	0.0	5.7	3.5	6.4	6.0	5.6	6.2
1	0.0	10.2	7.5	9.4	11.9	7.5	8.6	5.5

Group 13

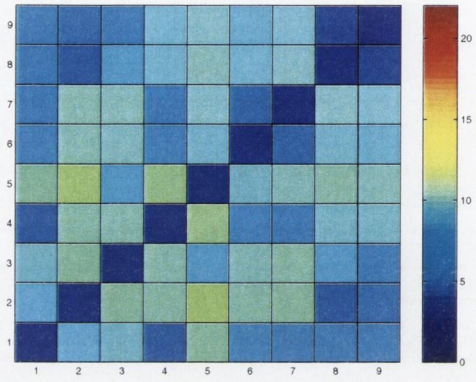


Figure E.80: Group 13 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
9	5	10	10	10	10	10
8	4	10	10	10	10	10
7	6	6	6	6	8	8
6	6	6	8	8	6	8
5	10	4	10	10	10	10
4	6	10	6	6	6	10
3	10	10	10	10	10	10
2	1	10	10	10	10	10
1	5	10	6	8	8	10

User	Euclidean Distances								
9	4.9	4.0	5.0	7.0	7.8	6.7	7.5	1.0	0.0
8	5.0	3.0	6.0	7.2	8.5	6.9	7.7	0.0	1.0
7	5.0	9.0	8.5	4.9	7.7	3.5	0.0	7.7	7.5
6	5.4	8.3	7.7	5.3	6.9	0.0	3.5	6.9	6.7
5	9.2	10.8	6.0	10.0	0.0	6.9	7.7	8.5	7.8
4	3.0	8.5	8.0	0.0	10.0	5.3	4.9	7.2	7.0
3	7.0	9.0	0.0	8.0	6.0	7.7	8.5	6.0	5.0
2	6.3	0.0	9.0	8.5	10.8	8.3	9.0	3.0	4.0
1	0.0	6.3	7.0	3.0	9.2	5.4	5.0	5.0	4.9

Group 14

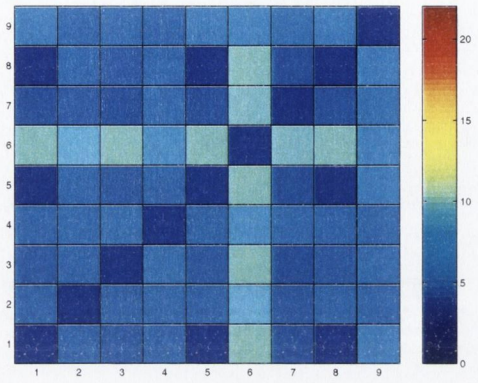


Figure E.81: Group 14 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
9	8	8	8	8	8	8
8	10	10	10	10	10	10
7	8	10	10	10	10	10
6	8	4	10	6	10	8
5	10	10	10	10	10	10
4	10	8	8	8	10	10
3	8	10	10	10	8	10
2	8	10	10	8	10	8
1	10	10	10	10	10	10

User	Euclidean Distances								
9	4.9	3.5	4.0	3.5	4.9	5.3	4.5	4.9	0.0
8	0.0	3.5	2.8	3.5	0.0	7.7	2.0	0.0	4.9
7	2.0	2.8	2.0	4.0	2.0	7.5	0.0	2.0	4.5
6	7.7	6.3	7.7	5.7	7.7	0.0	7.5	7.7	5.3
5	0.0	3.5	2.8	3.5	0.0	7.7	2.0	0.0	4.9
4	3.5	4.0	4.5	0.0	3.5	5.7	4.0	3.5	3.5
3	2.8	3.5	0.0	4.5	2.8	7.7	2.0	2.8	4.0
2	3.5	0.0	3.5	4.0	3.5	6.3	2.8	3.5	3.5
1	0.0	3.5	2.8	3.5	0.0	7.7	2.0	0.0	4.9

Group 15

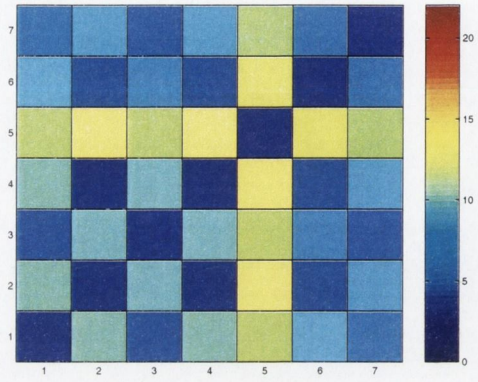


Figure E.82: Group 15 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
7	10	4	9	10	9	10
6	10	8	10	10	10	10
5	6	4	4	6	4	3
4	10	10	10	10	10	10
3	8	3	10	10	10	10
2	10	10	10	10	10	10
1	6	3	10	10	10	10

User	Euclidean Distances						
7	4.4	6.2	2.6	6.2	11.4	4.2	0.0
6	6.4	2.0	5.4	2.0	13.0	0.0	4.2
5	11.7	13.7	11.9	13.7	0.0	13.0	11.4
4	8.1	0.0	7.3	0.0	13.7	2.0	6.2
3	2.0	7.3	0.0	7.3	11.9	5.4	2.6
2	8.1	0.0	7.3	0.0	13.7	2.0	6.2
1	0.0	8.1	2.0	8.1	11.7	6.4	4.4

Group 16

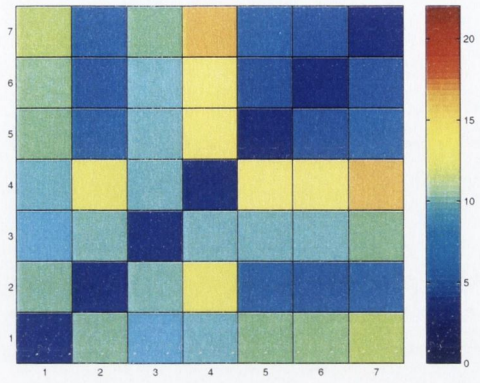


Figure E.83: Group 16 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
7	10	10	10	10	10	10
6	8	8	10	10	10	10
5	6	10	10	10	10	10
4	3	5	2	4	5	5
3	4	6	6	6	8	10
2	8	9	10	10	7	10
1	3	8	8	5	4	6

User	Euclidean Distances						
7	11.6	3.7	9.4	15.0	4.0	2.8	0.0
6	10.3	3.2	7.5	13.6	2.8	0.0	2.8
5	9.7	3.7	7.5	13.6	0.0	2.8	4.0
4	6.9	13.0	7.5	0.0	13.6	13.6	15.0
3	6.5	7.6	0.0	7.5	7.5	7.5	9.4
2	8.9	0.0	7.6	13.0	3.7	3.2	3.7
1	0.0	8.9	6.5	6.9	9.7	10.3	11.6

Group 17

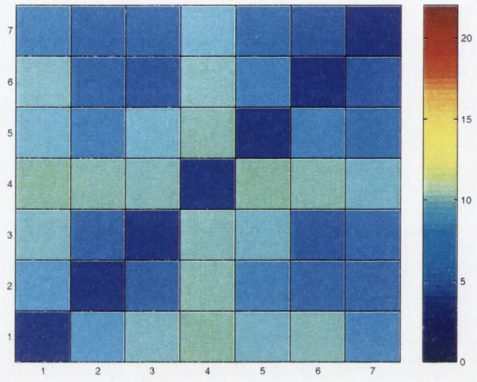


Figure E.84: Group 17 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
7	8	8	4	6	8	6
6	10	8	4	6	8	8
5	6	6	6	8	8	8
4	10	10	6	10	10	2
3	10	10	3	6	10	8
2	8	8	2	8	10	8
1	4	8	2	6	10	4

User	Euclidean Distances						
7	5.3	4.0	4.1	6.9	4.5	2.8	0.0
6	7.7	4.0	3.0	8.0	5.3	0.0	2.8
5	6.9	5.3	7.0	8.7	0.0	5.3	4.5
4	8.7	8.0	7.8	0.0	8.7	8.0	6.9
3	7.5	3.6	0.0	7.8	7.0	3.0	4.1
2	6.0	0.0	3.6	8.0	5.3	4.0	4.0
1	0.0	6.0	7.5	8.7	6.9	7.7	5.3

Group 18

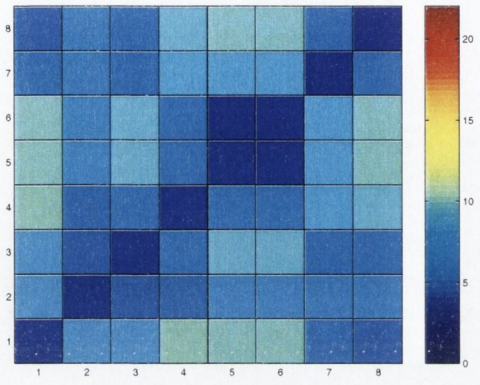


Figure E.85: Group 18 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
8	8	6	8	6	6	6
7	6	8	8	6	8	8
6	10	10	10	4	10	10
5	10	10	10	4	10	10
4	10	7	8	2	9	10
3	8	6	6	4	8	8
2	10	8	6	4	8	8
1	8	8	8	8	6	6

User	Euclidean Distances							
8	2.8	4.9	4.0	6.8	7.7	7.7	4.0	0.0
7	4.0	4.9	4.0	6.2	6.0	6.0	0.0	4.0
6	7.7	5.3	6.6	4.2	0.0	0.0	6.0	7.7
5	7.7	5.3	6.6	4.2	0.0	0.0	6.0	7.7
4	8.1	3.7	4.2	0.0	4.2	4.2	6.2	6.8
3	5.7	2.8	0.0	4.2	6.6	6.6	4.0	4.0
2	5.7	0.0	2.8	3.7	5.3	5.3	4.9	4.9
1	0.0	5.7	5.7	8.1	7.7	7.7	4.0	2.8

Group 19

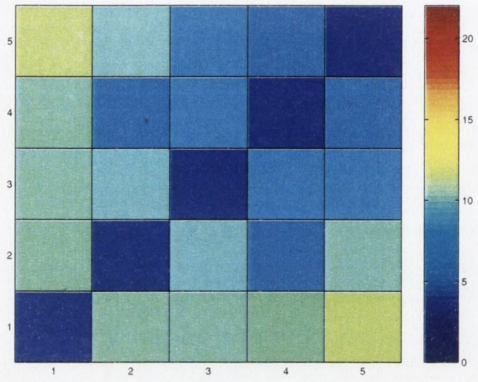


Figure E.86: Group 19 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
5	10	10	10	10	10	10
4	6	10	10	10	10	10
3	8	8	8	8	8	8
2	2	10	10	10	10	10
1	2	4	10	9	8	4

User	Euclidean Distances				
5	11.9	8.0	4.9	4.0	0.0
4	9.6	4.0	4.9	0.0	4.0
3	8.5	7.5	0.0	4.9	4.9
2	8.8	0.0	7.5	4.0	8.0
1	0.0	8.8	8.5	9.6	11.9

Group 20

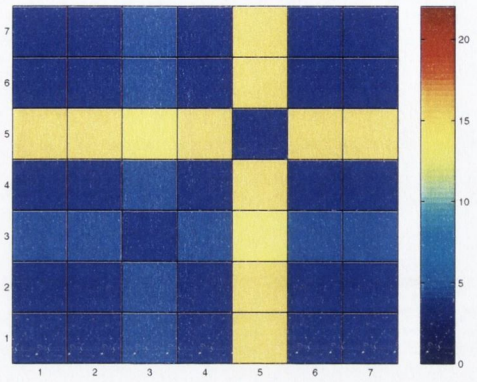


Figure E.87: Group 20 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
7	10	10	10	10	10	10
6	10	10	10	10	10	10
5	8	4	4	2	6	4
4	10	10	10	10	10	10
3	10	10	10	10	10	8
2	10	10	10	10	10	10
1	10	10	10	10	10	10

User	Euclidean Distances						
7	0.0	0.0	2.0	0.0	13.9	0.0	0.0
6	0.0	0.0	2.0	0.0	13.9	0.0	0.0
5	13.9	13.9	13.1	13.9	0.0	13.9	13.9
4	0.0	0.0	2.0	0.0	13.9	0.0	0.0
3	2.0	2.0	0.0	2.0	13.1	2.0	2.0
2	0.0	0.0	2.0	0.0	13.9	0.0	0.0
1	0.0	0.0	2.0	0.0	13.9	0.0	0.0

Group 21

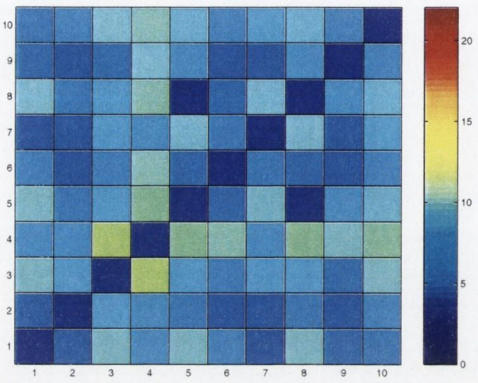


Figure E.88: Group 21 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
10	8	4	10	10	8	10
9	6	8	8	10	8	8
8	10	10	10	10	10	10
7	6	8	8	8	6	8
6	8	8	8	10	10	8
5	10	10	10	10	10	10
4	10	8	6	6	4	6
3	4	10	10	10	10	10
2	8	8	8	8	8	8
1	7	6	7	8	6	9

User	Euclidean Distances									
10	4.8	5.3	7.5	9.2	6.6	5.3	6.0	6.6	5.3	0.0
9	3.9	2.8	4.5	7.5	5.7	2.8	2.8	5.7	0.0	5.3
8	7.4	4.9	6.0	9.4	0.0	4.0	6.9	0.0	5.7	6.6
7	2.6	2.8	6.0	5.7	6.9	4.9	0.0	6.9	2.8	6.0
6	5.2	2.8	5.3	8.0	4.0	0.0	4.9	4.0	2.8	5.3
5	7.4	4.9	6.0	9.4	0.0	4.0	6.9	0.0	5.7	6.6
4	5.6	5.7	11.1	0.0	9.4	8.0	5.7	9.4	7.5	9.2
3	7.4	6.0	0.0	11.1	6.0	5.3	6.0	6.0	4.5	7.5
2	3.3	0.0	6.0	5.7	4.9	2.8	2.8	4.9	2.8	5.3
1	0.0	3.3	7.4	5.6	7.4	5.2	2.6	7.4	3.9	4.8

Group 22

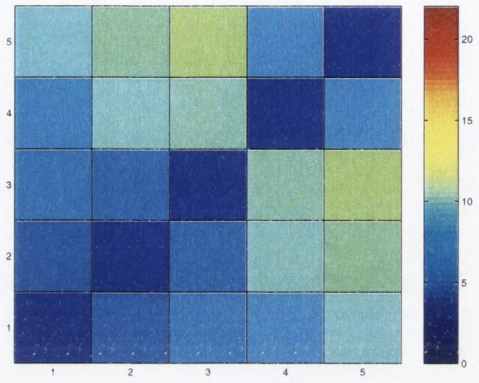


Figure E.89: Group 22 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
5	4	6	2	8	6	4
4	6	4	2	8	8	8
3	10	10	5	10	10	10
2	10	10	2	10	10	8
1	8	8	2	10	10	8

User	Euclidean Distances				
5	7.5	9.4	10.8	5.3	0.0
4	5.3	7.7	8.5	0.0	5.3
3	4.6	3.6	0.0	8.5	10.8
2	2.8	0.0	3.6	7.7	9.4
1	0.0	2.8	4.6	5.3	7.5

Group 23

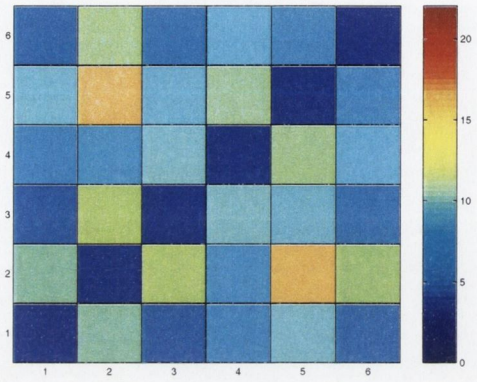


Figure E.90: Group 23 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
6	4	3	5	6	6	6
5	2	2	2	6	2	6
4	6	6	5	10	9	8
3	6	5	2	5	6	5
2	10	8	8	10	10	10
1	6	6	4	6	6	6

User	Euclidean Distances					
6	3.7	10.9	4.4	6.5	5.5	0.0
5	7.2	15.2	6.6	10.5	0.0	5.5
4	5.5	5.8	7.3	0.0	10.5	6.5
3	2.6	11.3	0.0	7.3	6.6	4.4
2	9.2	0.0	11.3	5.8	15.2	10.9
1	0.0	9.2	2.6	5.5	7.2	3.7

Group 24

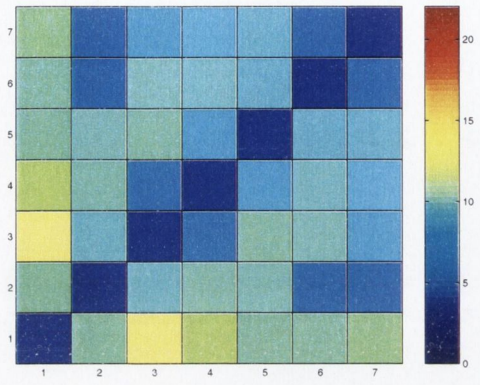


Figure E.91: Group 24 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
7	8	8	10	5	10	8
6	6	8	8	4	10	10
5	10	6	4	6	10	8
4	10	6	8	10	10	10
3	10	10	10	10	10	10
2	6	10	9	5	8	8
1	4	4	4	4	4	8

User	Euclidean Distances						
7	10.2	3.6	6.1	6.4	6.7	3.6	0.0
6	8.7	3.7	7.7	7.5	6.6	0.0	3.6
5	8.9	7.9	8.5	6.0	0.0	6.6	6.7
4	11.5	8.1	4.5	0.0	6.0	7.5	6.4
3	13.6	7.1	0.0	4.5	8.5	7.7	6.1
2	9.1	0.0	7.1	8.1	7.9	3.7	3.6
1	0.0	9.1	13.6	11.5	8.9	8.7	10.2

Group 25

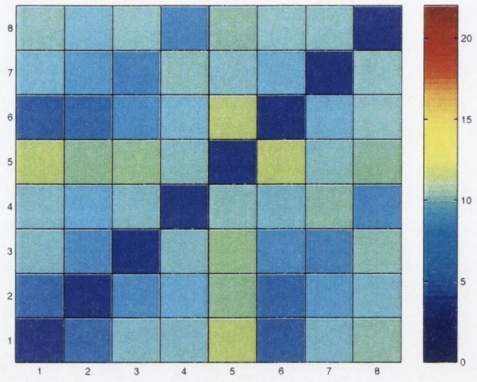


Figure E.92: Group 25 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
8	4	8	4	4	6	6
7	10	6	2	8	8	8
6	10	8	2	2	8	10
5	10	4	4	8	4	2
4	8	6	5	2	4	6
3	10	8	6	6	8	10
2	10	10	2	4	6	8
1	10	8	0	2	6	10

User	Euclidean Distances							
8	8.5	6.9	8.0	5.4	9.4	8.0	8.2	0.0
7	7.2	6.0	5.3	8.3	7.7	6.6	0.0	8.2
6	2.8	4.0	5.7	7.0	11.7	0.0	6.6	8.0
5	11.7	9.8	10.2	7.8	0.0	11.7	7.7	9.4
4	7.3	6.4	7.5	0.0	7.8	7.0	8.3	5.4
3	7.5	5.7	0.0	7.5	10.2	5.7	5.3	8.0
2	4.0	0.0	5.7	6.4	9.8	4.0	6.0	6.9
1	0.0	4.0	7.5	7.3	11.7	2.8	7.2	8.5

Group 26

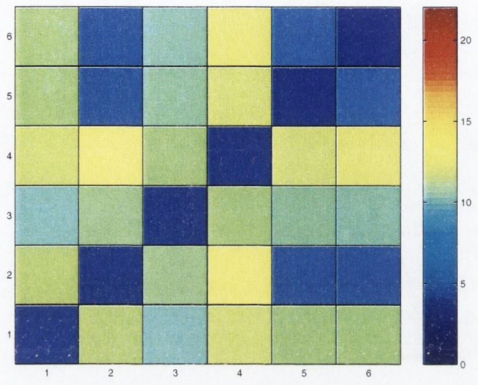


Figure E.93: Group 26 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
6	10	10	8	10	10	10
5	10	8	10	10	10	10
4	10	2	4	10	2	6
3	5	5	2	10	10	10
2	10	10	10	10	10	10
1	2	6	6	10	10	4

User	Euclidean Distances					
6	11.0	2.0	9.3	12.6	2.8	0.0
5	11.0	2.0	9.9	12.3	0.0	2.8
4	12.3	13.4	10.9	0.0	12.3	12.6
3	7.9	10.7	0.0	10.9	9.9	9.3
2	11.5	0.0	10.7	13.4	2.0	2.0
1	0.0	11.5	7.9	12.3	11.0	11.0

Group 27

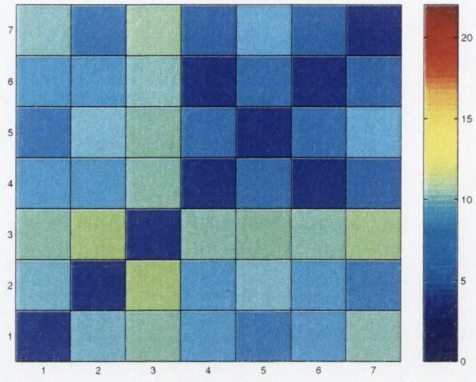


Figure E.94: Group 27 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
7	10	10	10	10	8	10
6	8	8	10	10	10	10
5	4	10	10	10	10	10
4	8	8	10	10	10	10
3	6	6	8	8	10	2
2	8	8	10	10	4	10
1	3	6	8	10	8	10

User	Euclidean Distances						
7	8.3	4.9	10.4	3.5	6.3	3.5	0.0
6	6.1	6.0	8.9	0.0	4.5	0.0	3.5
5	5.0	7.5	9.6	4.5	0.0	4.5	6.3
4	6.1	6.0	8.9	0.0	4.5	0.0	3.5
3	9.0	10.8	0.0	8.9	9.6	8.9	10.4
2	7.0	0.0	10.8	6.0	7.5	6.0	4.9
1	0.0	7.0	9.0	6.1	5.0	6.1	8.3

Group 28

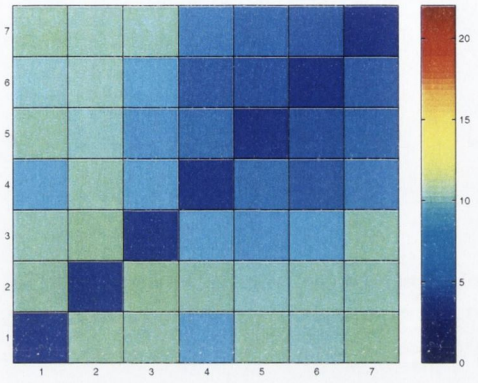


Figure E.95: Group 28 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
7	10	10	8	4	4	10
6	10	10	8	6	6	10
5	10	10	10	6	6	10
4	10	10	6	6	8	10
3	10	10	10	10	10	10
2	6	10	10	4	8	4
1	6	7	4	7	10	9

User	Euclidean Distances						
7	9.3	8.5	8.7	4.9	3.5	2.8	0.0
6	7.7	8.0	6.0	2.8	2.0	0.0	2.8
5	8.9	7.7	5.7	4.5	0.0	2.0	3.5
4	5.9	8.5	6.0	0.0	4.5	2.8	4.9
3	8.4	9.6	0.0	6.0	5.7	6.0	8.7
2	9.1	0.0	9.6	8.5	7.7	8.0	8.5
1	0.0	9.1	8.4	5.9	8.9	7.7	9.3

Group 29

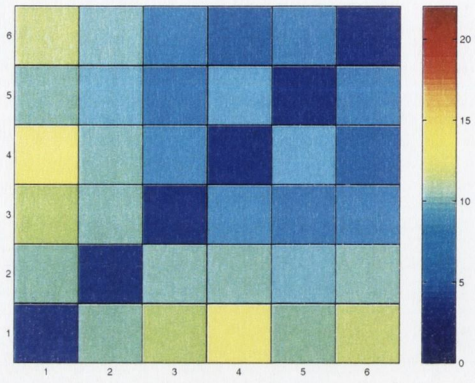


Figure E.96: Group 29 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
6	8	10	8	10	8	10
5	8	10	10	6	6	8
4	6	10	10	10	10	10
3	10	10	10	6	10	10
2	5	5	7	5	8	8
1	10	2	10	4	2	10

User	Euclidean Distances					
6	12.0	8.0	5.3	3.5	5.3	0.0
5	9.6	6.9	4.9	6.3	0.0	5.3
4	13.4	8.2	5.7	0.0	6.3	3.5
3	11.5	8.2	0.0	5.7	4.9	5.3
2	9.2	0.0	8.2	8.2	6.9	8.0
1	0.0	9.2	11.5	13.4	9.6	12.0

Group 30

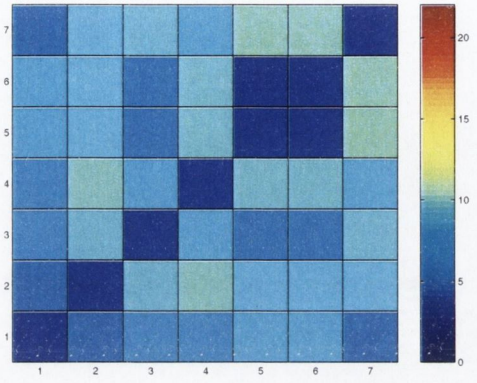


Figure E.97: Group 30 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
7	4	10	8	10	6	10
6	10	10	6	10	10	6
5	10	10	6	10	10	6
4	10	10	9	10	5	10
3	9	10	6	9	10	10
2	6	10	4	8	6	6
1	6	10	6	10	6	8

User	Euclidean Distances						
7	3.5	6.3	6.8	6.2	8.5	8.5	0.0
6	6.0	6.3	4.2	7.1	0.0	0.0	8.5
5	6.0	6.3	4.2	7.1	0.0	0.0	8.5
4	5.5	7.9	6.0	0.0	7.1	7.1	6.2
3	5.5	6.8	0.0	6.0	4.2	4.2	6.8
2	3.5	0.0	6.8	7.9	6.3	6.3	6.3
1	0.0	3.5	5.5	5.5	6.0	6.0	3.5

Group 31

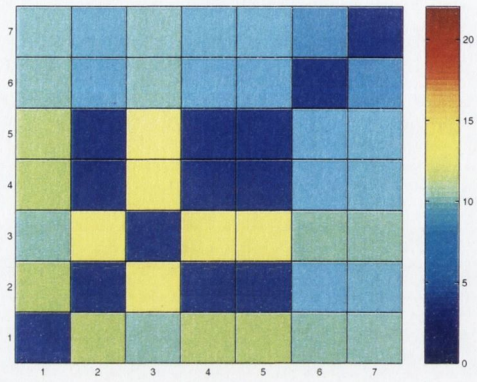


Figure E.98: Group 31 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
7	10	5	10	6	9	8
6	10	10	8	6	8	6
5	10	10	10	10	10	10
4	10	10	10	10	10	10
3	10	4	4	2	10	4
2	10	10	10	10	10	10
1	8	4	4	6	4	8

User	Euclidean Distances						
7	8.1	6.8	8.4	6.8	6.8	5.8	0.0
6	8.7	6.3	8.7	6.3	6.3	0.0	5.8
5	11.5	0.0	13.1	0.0	0.0	6.3	6.8
4	11.5	0.0	13.1	0.0	0.0	6.3	6.8
3	8.5	13.1	0.0	13.1	13.1	8.7	8.4
2	11.5	0.0	13.1	0.0	0.0	6.3	6.8
1	0.0	11.5	8.5	11.5	11.5	8.7	8.1

Group 32

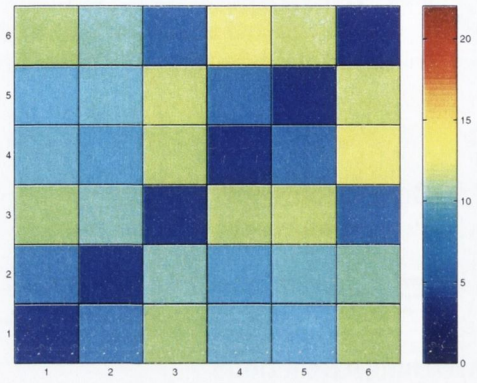


Figure E.99: Group 32 Euclidean Distance Matrix

User	Similarity Rating (from zero to ten)					
6	8	8	6	10	10	10
5	5	6	3	1	5	6
4	8	8	2	2	3	5
3	10	10	6	8	10	10
2	10	8	4	4	8	4
1	10	4	2	2	8	5

User	Euclidean Distances					
6	11.4	9.2	3.5	12.4	12.0	0.0
5	6.4	7.2	11.8	4.5	0.0	12.0
4	6.7	6.2	11.6	0.0	4.5	12.4
3	10.8	8.0	0.0	11.6	11.8	3.5
2	5.0	0.0	8.0	6.2	7.2	9.2
1	0.0	5.0	10.8	6.7	6.4	11.4

E.5 Story-by-Story Matrices

E.5.1 Evaluation text filenames

This list provides the key to the story-by-story matrices in sections E.5.2, E.5.3 and E.5.4..

1. 271102_AMDIraq8
2. 041202TM1745_AMDBudgetSources25
3. 041202TM1745_AMDBudgetSources14
4. 061202_AMDMurder4
5. 041202TM1513_AMDPyramid2
6. 271102_AMDFloods6
7. 041202TM1745_AMDBudgetReaction5
8. 221102_AMDflood9
9. 041202TM1513_AMDEclipse3
10. 041202TM1745_AMDMoriarty3
11. 051202_AMDMoriarty2
12. 041202TM1537_AMDTourism2
13. 041202TM1104AMDIraq5
14. 051202_AMDAnti-US3
15. 041202TM1745_AMDBudgetSources49
16. 271102_AMDIsrael8
17. 051202_AMDProdi7
18. 061202_AMDMacDoBomb2
19. 221102_AMDpisces5
20. 271102_AMDFloods4
21. 041202TM1745_AMDBudgetSources17
22. 221102_AMDflood8
23. 221102_AMDfirefighters8
24. 051202_AMDFlood2
25. 271102AMDFirebomb4

E.5.2 Story Pair Matrix

This matrix show the story pair number for each story combination. The key to the numbers of the matrix is set out in the list in appendix E.5.1.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1																									
2	161																								
3	124	73																							
4	185	41	123																						
5	131	158	179	52																					
6	265	98	114	145	187																				
7	46	148	189	165	230	136																			
8	188	149	104	121	194	178	271																		
9	22	18	224	217	143	35	268	128																	
10	112	172	192	4	222	138	206	198	176																
11	258	285	111	242	297	276	225	300	155	147															
12	139	42	59	66	287	201	157	33	68	278	246														
13	129	173	167	259	156	250	234	29	3	127	16	183													
14	229	212	87	162	82	286	51	119	209	232	295	60	153												
15	137	223	261	101	219	243	281	245	70	5	64	142	199	113											
16	177	7	275	289	38	203	49	151	54	25	24	81	50	56	55										
17	146	263	269	14	280	239	166	228	294	120	97	169	282	31	293	88									
18	241	84	292	6	214	266	15	171	244	21	19	108	44	2	37	109	125								
19	299	76	94	226	257	28	27	220	181	126	47	284	11	58	267	67	191	140							
20	160	291	174	77	168	251	159	9	45	238	237	105	115	254	231	279	164	130	200						
21	12	283	91	298	65	110	208	132	248	170	210	211	273	150	106	272	102	252	163	207					
22	154	253	255	117	86	43	72	63	78	190	99	75	118	8	204	10	213	36	182	61	39				
23	40	184	262	80	74	180	13	83	133	34	256	90	116	122	235	134	26	197	57	193	186	95			
24	32	135	270	205	288	62	277	202	227	216	247	175	196	17	195	236	260	96	79	30	100	218	69		
25	141	23	290	274	93	71	1	296	103	144	20	92	240	249	85	53	233	215	152	48	221	264	107	89	

E.5.3 Mean Ratings in Story-by-Story Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	YY																								
2	8.57	YY																							
3	9.00	5.70	YY																						
4	8.71	8.86	8.50	YY																					
5	8.80	7.50	7.57	7.78	YY																				
6	X	8.29	8.40	9.00	8.71	YY																			
7	8.89	4.50	3.83	8.00	X	7.17	YY																		
8	7.33	6.25	7.89	7.70	X	9.57	X	YY																	
9	7.75	9.78	X	X	8.86	6.33	X	7.60	YY																
10	9.40	6.83	6.67	9.33	X	6.83	X	X	10	YY															
11	X	X	9.60	X	X	X	X	X	8.67	3.13	YY														
12	7.71	8.14	6.50	8.63	X	X	6.63	7.67	8.00	X	X	YY													
13	2.60	7.33	7.43	X	8.33	X	X	7.63	9.50	7.60	8.11	8.00	YY												
14	X	X	9.00	8.86	8.89	X	7.22	9.43	X	X	X	7.63	6.67	YY											
15	6.50	X	X	9.14	X	X	X	X	8.75	8.17	7.63	6.29	X	9.20	YY										
16	6.43	9.70	X	X	8.67	X	8.67	8.14	9.56	8.22	8.88	9.56	7.11	8.00	8.25	YY									
17	7.33	X	X	9.50	X	X	6.14	X	X	8.86	8.00	7.83	X	7.33	X	9.43	YY								
18	X	9.33	X	5.50	X	X	9.44	9.17	X	8.63	9.25	8.25	8.89	8.71	8.56	5.60	8.00	YY							
19	X	8.67	7.86	X	X	7.50	7.25	X	9.75	8.70	8.67	X	9.40	8.38	X	7.67	7.33	7.43	YY						
20	9.71	X	9.33	8.67	9.00	X	9.43	8.80	8.11	X	X	8.00	9.67	X	X	X	9.13	9.20	X	YY					
21	9.60	X	6.00	X	8.25	8.40	X	7.60	X	7.83	X	X	X	7.50	4.75	X	6.29	X	8.67	X	YY				
22	8.33	X	X	9.25	6.00	9.56	4.63	4.63	9.56	4.50	3.86	8.44	8.50	8.90	X	9.40	X	8.83	7.57	9.22	7.75	YY			
23	6.88	7.14	X	8.89	8.44	8.00	7.30	9.56	6.00	7.33	X	9.00	8.67	8.00	X	4.86	7.75	X	7.75	6.67	8.00	7.71	YY		
24	6.67	4.33	X	X	X	8.63	X	X	X	X	X	8.13	X	8.89	X	X	X	8.71	8.89	7.88	7.14	X	5.13	YY	
25	7.57	8.88	X	X	8.00	7.63	9.43	X	8.89	8.86	8.25	8.00	X	X	8.89	5.33	X	X	7.29	9.11	X	X	8.13	9.00	YY

E.5.4 Mode Ratings in Story-by-Story Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	YY																								
2	10	YY																							
3	10	4	YY																						
4	10	10	8	YY																					
5	10	8	10	10	YY																				
6	X	8	10	10	10	YY																			
7	8	2	2	10	X	6	YY																		
8	8	8	8	8	X	10	X	YY																	
9	10	10	X	X	10	4	X	10	YY																
10	10	6	4	10	X	6	X	X	10	YY															
11	X	X	10	X	X	X	X	X	10	2	YY														
12	10	8	8	10	X	X	8	8	8	X	X	YY													
13	2	8	8	X	10	X	X	8	10	10	8	10	YY												
14	X	X	10	8	10	X	8	10	X	X	X	6	10	YY											
15	6	X	X	10	X	X	X	X	10	10	6	4	X	10	YY										
16	6	10	X	X	8	X	10	10	10	10	10	10	6	8	8	YY									
17	8	X	X	10	X	X	6	X	X	10	10	8	X	8	X	10	YY								
18	X	10	X	6	X	X	10	10	X	10	10	8	6	10	10	2	10	YY							
19	X	10	10	X	X	6	8	X	10	8	8	X	10	10	X	8	10	8	YY						
20	10	X	10	10	10	X	10	10	8	X	X	8	10	X	X	X	10	10	X	YY					
21	10	X	3	X	10	10	X	8	X	10	X	X	X	10	4	X	8	X	10	X	YY				
22	10	X	X	10	3	10	4	2	10	2	2	10	10	10	X	10	X	8	10	10	10	YY			
23	8	10	X	10	10	10	8	10	6	10	X	10	10	8	X	6	10	X	8	6	10	10	YY		
24	6	4	X	X	X	10	X	X	X	X	X	10	X	10	X	X	X	10	8	10	6	X	4	YY	
25	8	10	X	X	10	6	10	X	10	8	10	8	X	X	10	4	X	X	10	10	X	X	8	10	YY