Manchester Metropolitan University

e-space

Manchester Metropolitan University's Research Repository

Please cite the Published Version

O'Shea, James (2010) A framework for applying short text semantic similarity in goal-oriented conversational agents. Doctoral thesis (PhD), Manchester Metropolitan University.

Downloaded from: https://e-space.mmu.ac.uk/624481/

Usage rights: © In Copyright

Enquiries:
If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines)

A FRAMEWORK FOR APPLYING SHORT TEXT SEMANTIC SIMILARITY IN GOAL-ORIENTED CONVERSATIONAL AGENTS

JAMES D. O'SHEA

A thesis submitted in partial fulfilment of the requirements of the Manchester Metropolitan University for the degree of Doctor of Philosophy

July 2010

Department of Computing and Mathematics
The Manchester Metropolitan University

Copyright

Copyright in text of this thesis rests with the author. Copies (by any process) either in full, or extracts, may be made only in accordance with instructions given by the Author and lodged in the Manchester Metropolitan University Library. Details may be obtained from the Librarian. This page must form part of any such copies made. Further copies (by any process) of copies made in accordance with such instructions may not be made without permission of the Author.

The Ownership of any intellectual property rights which may be described in this thesis is vested in the Manchester Metropolitan University, subject to any prior agreement to the contrary, and may not be made available for use by any third party without the written permission of the University, which will prescribe the terms and conditions of any such agreement.

Further information on the conditions under which disclosures and exploitation may take place is available from the Head of Department of Computing and Mathematics.

Abstract

Existing Conversational Agents (CAs) have several disadvantages. The most serious is that the CAs that humans find most coherent and intelligent are based on the pattern matching technique, which is labour intensive and results in CAs that are difficult to maintain. The main alternative technique, Natural Language Processing, produces CAs which have a high computational complexity and are unlikely to scale well when used by large numbers of people. These limitations have prevented CAs from realising their huge potential in practical applications.

This thesis concerns a framework for the development of a new generation of CAs. The key component is Short Text Semantic Similarity (STSS). Replacing pattern matching rules by measurement of the similarity between user utterances and prototype statements results in CAs which are simple to develop and maintain, and are also computationally efficient.

STSS algorithms are a recent development and a method is required to evaluate and compare the stream of new emerging algorithms before they are incorporated into CAs. This thesis investigated the development of benchmark datasets for the evaluation of such new algorithms.

A second strand of work concerned the development of a new model of STSS, taking account of Dialogue Acts and Valence, two factors which have not been considered in previous models.

The benefits and achievements of this work include identification of the best methodology for obtaining ground truth similarity from human raters, the production of two gold standard benchmark datasets for evaluation of STSS measures, the proposal of a factor-based model of STSS and the development of a set of computationally efficient classifiers for the question dialogue act.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other University or other institution of learning.

Dedication

To my family

Acknowledgements

I would like to express my deepest gratitude to my supervisors Dr Zuhair Bandar and Dr Keeley Crockett for their guidance, support and patience during the course of my research.
I would like to thank the heads of department during the period of my work, in particular for allowing the use of personal development funds to support the compensation of participants in my experiments.
I would also like to thank the large number of anonymous participants who took part in the experiments to create the datasets used in this work.
Special thanks go to my parents, family and friends for their encouragement and support throughout this work.

"In physical science the first essential step in the direction of learning any subject is to find principles of numerical reckoning and practicable methods for measuring some quality connected with it. I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be."

Lord Kelvin in Popular Lectures and Addresses (PLA), vol. 1, "Electrical Units of Measurement", 1883-05-03,

Contents

Chapter 1 Introduction

1.1 Contribution 1
1.2 Conversational Agents..... 2
1.2.1 Pattern Matching Chatterbots 2
1.2.1.1 Eliza..... 4
1.2.1.2 Racter..... 6
1.2.1.3 Sylvie..... 7
1.2.1.4 Alice..... 9
1.2.1.5 Summary of Pattern Matching Chatterbots..... 10
1.2.2 NLP based systems..... 11
1.2.3 Goal-Oriented Conversational Agents..... 13
1.2.4 Fundamental problems of current CA approaches..... 16
1.2.5 The proposed solution – Short Text Semantic Similarity..... 17
1.2.5.1 Short texts – working definition..... 17
1.2.5.2 Contribution of STSS to future CAs..... 18
1.3 Related research in the field of similarity..... 19
1.3.1 Definitions of Similarity and Semantic Similarity..... 19
1.3.1.1 Definition of Semantic Similarity..... 20
1.3.2 Prior work on Similarity..... 21
1.3.2.1 Fundamental work by Tversky..... 21
1.3.2.2 Similarity in Cognitive Science, Neuropsychology and Neuroscience..... 23
1.3.2.3 Similarity in other fields..... 33
1.3.3 Real-world acceptance of similarity..... 35
1.3.3.1 Visual similarity and its practical applications..... 35
1.3.3.2 Practical applications of Semantic Similarity in NLP and Information Retrieval..... 37
1.3.3.3 Question Answering and Information Retrieval..... 37
1.3.3.4 Copyright and Plagiarism..... 39
1.3.3.5 Word Sense Disambiguation..... 40
1.3.3.6 Data Cleansing and Ontology Alignment..... 43
1.3.3.7 The Semantic Web..... 44
1.3.3.8 Other work involving the Web and Similarity..... 46
1.4 Key Concepts..... 48
1.4.1 Short texts or sentences? 48
1.4.2 Similarity vs. Relatedness, Antonymy and kinds of similarity... 49
1.4.3. Origins of Semantic Similarity measures..... 50
1.4.3.1 Euclidian Distance..... 50
1.4.3.2 The Vector Space Model..... 50
1.4.4 Fundamental Models of Short Text Semantic Similarity, STASIS and LSA..... 52
1.4.4.1 STASIS..... 52
1.4.4.2 LSA..... 56
1.4.5 Evaluation: Agents vs. Algorithms..... 59
1.4.5.1 Holistic Evaluation of Agents..... 63
1.4.5.2 Aspirational Subjective Measures..... 63
1.4.5.3 Attempted Subjective Measures..... 63
1.4.5.4 Objective measures..... 65

1.4.5.5. Evaluation of algorithms 66
1.4.5.6 Evaluating STSS measures 66
1.5 Conclusions 67
1.5.1 Relationships between chapters 69

Chapter 2 Production of STSS-65, the first Short Text Semantic Similarity dataset

2.1 Contributions to research in the field 72
2.2 Obstacles to the creation of an STSS dataset 73
2.2.1 The NLP Literature set 74
2.2.2 The Paired Dictionary Set 74
2.3 Cutting the Gordian knot – the STSS-65 dataset 77
2.4 Requirements of a Gold Standard dataset 78
2.4.1 Representative sampling 79
2.4.2 Precision and Accuracy 79
2.4.3 Measurement scale 79
2.5 The Rubenstein & Goodenough Legacy 80
2.5.1 Method of presentation of materials 80
2.5.2 Method of randomization of materials 81
2.5.3 Instructions, guidance and measurement scale definition 82
2.5.4 Sampling the population for participants 83
2.5.5 Summary of work after Rubenstein & Goodenough 83
2.6 Creation of the STSS-65 dataset 84
2.6.1 Selection of materials 84
2.6.2 Similarity measurement scales 86
2.6.2.1 Likert-like scales 86
2.6.2.2 Thurstone scales 88
2.6.2.3 Selection of a scale technique 89
2.6.3 Population and Sampling 90
2.6.4 Randomization 91
2.6.5 Procedures and Instructions 91
2.6.6 Results 93
2.7 Discussion of the STSS-65 dataset 102
2.7.1 Overall consistency between STSS-65 and Rubenstein & Goodenough 102
2.7.2 The challenge of medium similarity 104
2.8 Validation of the STSS-65 dataset 106
2.8.1 Common Instructions and variations 107
2.8.2 Population and Sampling 108
2.8.3 Age 108
2.8.4 Education 108
2.9 Experimental factors 109
2.9.1 Order 109
2.9.2 Anchor 109
2.9.3 Requirements for valid ANOVA analysis 109
2.9.3.1 Test for normal distribution 110
2.9.3.2 Test for equal variances 111
2.9.4 ANOVA Analysis of Order and Anchor 112
2.10 Analysis of results 113
2.10.1 Effect of varying Order and Anchor 113

2.10.2 Correlation between experiments..... 116
2.10.3 Consistency of judgment within experiments..... 116
2.10.4 The effect of the factors on noise..... 117
2.10.5 Effect of order of presentation of sentences within a pair..... 117
2.11 Applications and conclusions..... 118
2.11.1 Application to recently developed measures..... 118
2.11.2 Conclusions..... 119

Chapter 3 STSS-131 - the new benchmark dataset for STSS

3.1 Introduction..... 121
3.2 Creation of new benchmark dataset, STSS-131..... 121
3.2.1 Selecting stimulus words..... 122
3.2.1.1 Choosing the size of the stimulus word set..... 122
3.2.1.2 Decomposing the words of the English language into a hierarchy of classes..... 123
3.2.1.3 Mapping the classes from the decomposition onto a sampling frame..... 148
3.2.1.4 Populating the slots in the frame with words using an Opportunistic Random Search..... 148
3.2.2 Experiment 3.1 Collection of the sentence pairs..... 154
3.2.2.1 Creating a blocked design..... 154
3.2.2.2 Sampling the population..... 155
3.2.2.3 Design of a questionnaire instrument and instructions..... 156
3.2.2.4 Conducting the pilot study..... 158
3.2.2.5 Conducting the experimental trials..... 159
3.2.3 Selection of the 64 sentence pairs..... 160
3.2.3.1 Creation of the database..... 161
3.2.3.2 Materials selection by committee 162
3.2.4 Experiment 3.2 Collecting similarity ratings for the sentence pairs..... 162
3.2.4.1 Pilot study and modifications to materials..... 163
3.2.4.2 Specification, selection and recruitment of a suitable population sample..... 164
3.2.4.3 Experimental trials..... 165
3.2.4.4 Results of discriminability experiment..... 166
3.2.4.5 Analysis of results..... 173
3.2.4.6 Application of STSS-131: Comparing the performance of two different STSS measures..... 180
3.3 The Discriminability Study..... 183
3.3.1 Experiment 3.3 creation of a set of paragraph contexts..... 185
3.3.1.1 Materials - Selecting a sample of sentence pairs..... 185
3.3.1.2 Creating a blocked design..... 186
3.3.1.3 Sampling the population..... 187
3.3.1.4 Design of a questionnaire data collection instrument..... 188
3.3.1.5 Paragraph data collection trials..... 189
3.3.2 Experiment 3.4 Collecting discriminability ratings for the sentence pairs..... 189
3.3.2.1 Extraction of contextual paragraphs..... 189
3.3.2.2 Pilot study of the discriminability materials..... 191

3.3.2.3 Blocked design of experiments to collect discriminability ratings 192
3.3.2.4 Design of questionnaire instrument to collect sortings for discriminability measurement 193
3.3.2.5 Specification, selection and recruitment of a suitable population sample 195
3.3.2.6 Experimental trials 195
3.3.2.7 Results of discriminability experiment 195
3.4 Discussion 204
3.4.1 Distinguishing features of STSS-131 204
3.4.2 The discriminability study 205
3.4.3 Motivation for discriminability study 205
3.4.4 Comparison of STASIS and LSA using STSS-131 206
3.4.5 Weaknesses of current STSS measures 208

Chapter 4 Towards a New Model of Short Text Semantic Similarity

4.1. Introduction 209
4.2 The Dialogue Acts problem 209
4.2.1 Definition of a Dialogue Act 209
4.2.2 Dialogue Acts and STSS measurement 211
4.3 The Valence Problem 212
4.3.1 Definition of Valence 212
4.3.2 Valence and STSS measurement 213
4.4 A Factor-based model of Short Text Semantic Similarity 214
4.4.1 The proposed model 214
4.4.2 Requirements for developing the proposed model 215
4.5 Development of the Slim Function Word DA classifier 215
4.5.1 The Classification of Dialogue Acts 216
4.5.1.1 Dialogue Act Taxonomies 216
4.5.1.2 DA Feature extraction 218
4.5.1.3 Dialogue Act Classifiers 224
4.5.2 Selection of a classifier for Dialogue Acts 225
4.6 Decision Trees 226
4.6.1 Decision Tree construction 227
4.6.2 Decision Tree pruning 229
4.6.3 Coping with missing attribute values 231
4.6.4 Windowing 232
4.6.5 Grouped Attribute Values 232
4.6.6 Cross Validation 232
4.6.7 Illustrative Examples of Decision Tree Processes 233
4.7 The proposed Slim Function Word Classifier algorithm 237
4.7.1 Analysis of questions and non-questions 238
4.7.1.1 Grammatical classification 239
4.7.1.2 Domain-based classification 240
4.7.2 Forms of questions used in this study 242
4.7.2.1 Straightforward questions 242
4.7.2.2 Straightforward questions with preambles 242
4.7.2.3 Simulated clauses 243
4.7.2.4 Omitted question classes 244
4.7.2.5 Straightforward vs. Difficult non-questions 245

4.7.3 Formulation of the general hypothesis for the experiments in this chapter 246
4.7.4 Collection of raw question / non-question data 247
4.7.5 Creating datasets for particular combinations of the question / non-question forms 250
4.7.6 Preprocessing the datasets for feature extraction 251
4.7.6.1 Removing punctuation and spacing 252
4.7.6.2 Expanding contractions 252
4.7.6.3 Extending to a fixed length record 253
4.7.6.4 Feature extraction 253
4.7.6.5 Token scale type 255
4.8 Dataset creation 256
4.8.1 Preambles 256
4.8.2 Simulated clauses 256
4.9 Specific hypotheses 257
4.10 Experimental parameters 258
4.10.1 Baseline measures and statistical significance 258
4.10.2. N-fold Cross Validation 259
4.10.3 Pruning 260
4.11 Results of Experiments 261
4.12 Comparative analysis of experiments 279
4.13 General Conclusions 283

Chapter 5 Optimising the structure of function word features

5.1 Introduction 285
5.2 Function word taxonomy 286
5.3 Top-level decomposition 287
5.4 Decomposition of the major categories 290
5.4.1 Decomposition of the pronouns 290
5.4.2 Decomposition of the determiners 291
5.4.3 Decomposition of the link words 294
5.4.4 Auxiliaries 294
5.5 Application of the taxonomy in the new feature encoding method 294
5.5.1 Clustered coding 296
5.5.2 Grouped coding 296
5.6 Experimental Methodology 296
5.6.1. Selection of experiments 297
5.7 Results of experiments using Clustered Coding 298
5.7.1. Summary of Clustered Function Words Experiments 307
5.8 Results of experiments using Grouped Coding 308
5.8.1 Summary of Grouped Function Words Experiments 318
5.9 Supplementary Experiments 319
5.9.1 Summary of mixed Simple /Difficult Non-questions 324
5.10 Overview of experimental results 325
5.11 Comparative analysis of experiments 326
5.11.1 The outcome of clustering function words 326
5.11.2 The outcome of grouping function words 327
5.11.3 Comparison of clustered vs. grouped function words 328
5.12 Discussion 329
5.12.1 Analysis of the original intuitions 329

5.12.2 Possible explanations for the relative performances of clustered vs. grouped features 330
5.12.2.1 Duplicated records 330
5.12.2.2 Effect of restrictions on partitioning 330
5.13 Chapter Conclusions 332

Chapter 6 Discussion, Conclusions and Future Work

6.1 Discussion 333
6.2 Conclusions 335
6.3 Future work 337

References 341

Appendices 361

Appendix 1

1.1: Dialogue Samples from NLP-based systems

Appendix 2

2.1 Ethics statement (for Experiments 1, 2.1 and 2.4)
2.2 Instructions (for Experiments 1, 2.1 and 2.4)
2.3 Sample question sheet (for experiments 1 and 2.1)
2.4 Personal Data Sheet (for Experiments 1, 2.1, 2.2, 2.3, 2.4)
2.5 Ethics statement (for Experiments 2.2 and 2.3)
2.6 Instructions (for Experiments 2.2 and 2.3)
2.7 Sample card (for Experiments 2.2 and 2.3)
2.8 Sample rating sheet (without semantic anchors for Experiment 2.2)
2.9 Sample rating sheet (with semantic anchors for Experiment 2.3)
2.10 Sample instruction sheet (without semantic anchors for experiment 2.4)

Appendix 3

3.1 Blocked design for distribution to materials to participants (Experiment 3.1)
3.2 Randomised word blocks for experiment 3.1
3.3 Sample flier/poster used to recruit participants
3.4 Ethics statement / covering letter for Experiment 3.1 (sentence capture)
3.5 Experiment 3.1 instructions
3.6 Experiment 3.1 extracts from sentence capture form
3.7 Personal data sheet(for experiment 3.1)
3.8 Dialogue Act clarification sheet (for experiment 3.1)
3.9 Themes investigated for experiment 3.1
3.10 Word and Theme Supplements Experiment 3.1
3.11Sentences capture by Experiment 3.1
3.12 Pilot study version, call for volunteers (with link to ethics statement on web)
3.13 Sentences selected for paraphrase generation
3.14 Example paraphrased sentence capture sheet (experiment 3.1 modifications to materials)

3.15 Paraphrased sentences for high similarity pairs (experiment 3,1 modifications to materials)
3.16 Experiment 3.2 Selected sentence pairs with predicted similarity bands and selection criteria
3.17 sample sentence pair card
3.18 A sample of the comparison of students and general population
3.19 Ethics statement / covering letter for Experiment (paragraph capture)
3.20 Experiment 3.3 extracts from pararagraph capture form
3.21 A sample of the paragraphs selected for the blocked design experiment
3. 23 Extra information for unsupervised participants (experiment 3.4)
3.24 Block design for Experiment 3.4
3.25 Instructions for Experiment 3.4 (discriminability rating)
3.26 Extract from discriminability rating form

Appendix 4

4.1 Contraction list lookup table
4.2 Function word table used for tokenising sentences
4.3 A sample of the 600 Straightforward Questions from the dataset
4.4 A sample of the 600 Straightforward Non-questions
4.5 A sample of the 600 Difficult Nonquestions
4.6 1-word preambles
4.7 2-word preambles
4.8 3-word preambles

Appendix 5

5.1 Clustered Function Word Table
5.2 Grouped Function Word Table

Publications arising from the work

(Regulation 43.9: "Copies of relevant published material should either be bound in with the thesis or placed in an adequately secured pocket at the end of the thesis.")
N.B. These papers are not included in the electronic version of the thesis for copyright reasons. Information on accessibility can be found at the Semantic Similarity resource site http://semanticsimilarity.wordpress.com/

Chapter 1 Introduction

1.1 Contribution

This chapter defines the work performed, explaining the motivation for it, providing the historical background of research in the field and reviewing some key concepts that permeate the following chapters. This work concerns the development of key components of a framework for a new generation of Conversational Agents (CAs).
The term "Conversational Agent" is interpreted in different ways by different researchers; however the essence of CAs is natural language dialogue between the human and an application running on a computer. This thesis also uses the term "Goal-Oriented Conversational Agent", to distinguish between CAs that aspire to interact with a practical computer application and those which have an architecture which enables them to achieve a goal via the dialogue. Therefore Goal-Oriented Conversational Agents are a subset of the general class of CAs.
The main contribution of the work falls into three major areas:

- The first is the production of two benchmark datasets for evaluating the proposed new generation of CAs. These datasets are the first of their kind and are expected to be adopted as the Gold Standard for evaluating this class of CA.
- The second contribution is the establishment of a scientifically rigorous methodology for other researchers to create short text similarity datasets which will not only be useful in the evaluation of CAs but could also be used by the wider fields of dialogue management, information retrieval and cognitive science. This approach will define the processes for creating data items, collecting human ratings and statistical analysis of results. This methodology is designed to facilitate expansion of the dataset by pooling new materials from different researchers without fear of confounding factors.
- The third contribution is the development of a factor-based model of Short Text Semantic Similarity. This uses three factors, the Dialogue Act (DA), Propositional Content and Valence. This model is expected to contribute to increased accuracy of natural language understanding in CAs.

Empirically investigating the feasibility of such a model requires a method of identifying the dialogue act of a sentence. Furthermore, to fit the goal of deploying such models in scalable CAs requires the model to be computationally efficient.

To achieve the goal of computational efficiency, a method has been devised which is not only novel but runs counter to mainstream thinking in the field, by using only function words. Until now, function words have been considered to make no useful contribution to semantic similarity measurement and, apart from question words (who, what, where etc.) have little place in DA classification.

This chapter provides a historical background and review of Chatterbots & CAs, Similarity Measures, Neuroscience and the evaluation of CAs.

The central plank of this thesis is the proposal for a new approach to producing CAs using similarity algorithms. Similarity is an attribute extensively studied in diverse fields such as psychology, neuroscience and Natural Language Processing (NLP). Consequently, this chapter goes on to provide a thorough background of the concept of similarity. The coverage of neuroscience is particularly relevant as it removes a major stumbling block to progress in the development of Short Text Similarity datasets affecting the work in chapter 3.

1.2 Conversational Agents

The field of CAs has a pedigree almost as old as the computer itself (under various guises). It can be traced back to the work of Alan Turing where dialogue was proposed as a means of testing early concepts of machine intelligence (Turing, 1948, Turing, 1950, Turing, 1952). Although researchers tend to use some of the terminology interchangeably, in this thesis the general term Conversational Agent is used to cover a variety of approaches, which have been categorised as:

- Chatterbot-based systems
- NLP-based Dialogue Management Systems
- Goal-Oriented CAs

An analysis of the three approaches with respect to the production of scalable CAs will show that there are challenging obstacles to producing systems that can interact with users in real-time.

1.2.1 Pattern Matching Chatterbots

Chatterbots were the earliest form of CAs. They are programs that engage a human in social conversation, with the general intention of keeping the human engaged in the conversation for as long as possible. Much of the interest in developing chatterbots was

aimed at creating a machine that could pass the original Turing Test, which requires a computer to pass as human after 5 to 10 minutes of conversation.

The Eliza chatterbot (Weizenbaum, 1966) is an enduring example from the 1960s, although ironically it was designed to invalidate the Turing Test by creating a simplistic machine that obviously couldn't think but could still engage humans in convincing conversation. Most chatterbot development is small-scale and associated with web communities with a level of hobbyist support (Aimless, Conversive, Wallace, 2009).

Pattern matching mechanisms have been developed since the 1960s as the underlying technology of Chatterbots and subsequently CAs. Pattern matching has been identified as one of the most common and capable methods for developing dialogues that seem to be coherent and intelligent to users (Bickmore and Giorgino, 2006).

In pattern matching, a user utterance is scored against sets of patterns containing string text fragments and wildcards to find the closest match. Patterns are grouped within rules which in turn are contained in a script file (Sammut, 2001). When a script is executed, the user utterance is compared to the patterns and the closest match results in the relevant rule firing. This generates a suitable response to the user. For a chatterbot that is generally the end of the matter (although some variables such as the user's name may be set for later use in the conversation) - but they have no fundamental mechanisms to support a conversation that seeks to achieve a goal.

An application may require many script files covering the various conversational contexts (Sammut, 2001) that may occur and there are also mechanisms for switching contexts (this is analogous to the modularisation of code in conventional programming). Creating scripts is a highly skilled craft (Michie, 2001) requiring the anticipation of user utterances, generation of permutations of the utterances and generalization of patterns through the replacement of selected terms by wild cards. Modifications to rules containing the patterns can impact on the performance of other rules and modern pattern matching systems contain many parameters that further modify their behaviour.

The main strengths of pattern matching are:

- It works well within its limits and it's about the only technique, currently, that works at all for extended dialogues. Given a single user utterance a pattern matching system will often fire the correct rule and return an appropriate response.
- Pattern matching systems have good potential for scalability to support large numbers of users in real-time. They do not require complex pre-processing stages such as stemming, pos-tagging, syntactical repair and parsing. A real-world system

could require millions of instances of the CA to be running simultaneously on the organisation's servers.

- The computational engines for pattern matching systems are well-developed and robust; tested and deployed systems are rarely crashed by unexpected user input.
- Pattern matching separates out language skills from coding skills. People with language skills can become scripters without learning a great deal of computer science.

However, pattern matching also suffers from a number of weaknesses, some of which are:

- Writing patterns which match user inputs effectively is a labour intensive process and the scripters must be highly skilled at selecting key words or phrases and integrating them with wildcards.
- The CA's responses to the user must also be crafted to maintain the conversation along predictable lines. Transactions which are plausible in isolation can be stilted or incoherent as a complete conversation.
- The organisation of rules into coherent contexts involves another set of skills, similar to the design of coherent modules in conventional programming. Failure to do this results in systems that are difficult to test and debug. Such systems could also easily be destabilised by the addition of a single rule.

These drawbacks have an impact on development costs, maintainability and scalability. Consequently real-world implementations have failed to prosper.
Milestones in chatterbot development include Eliza, created with the aim of collecting evidence to counter Turing's proposals, Racter, Sylvie and Alice. The producers of Sylvie, Verbots, now operate commercially as Conversive (Conversive) but do not appear to be active in terms of demonstrator sites or client lists. Because pattern matching represents the best attempt so far to produce CAs, these systems will be described in further detail.

1.2.1.1 Eliza

Eliza was created by Joseph Weizenbaum (Weizenbaum, 1966) as a tool to explore natural language conversation between humans and computers. Its best-known implementation is as a Rogerian therapist. Eliza uses one keyword and structural features from the user utterance to select a decomposition / recomposition rule pair and formulate a reply. Keywords can be prioritised so that the most significant is used when several occur in the utterance. Recomposition is a very interesting idea and there has been little work since on

generating responses as opposed to simply reeling them off from memory. For example, answers returned by Question Answering systems can be as crude as a list of snippets extracted from the web or a specific ontology (Smucker et al., 2007, Marton, 2006).

One of the features used to create a sense of realism in Eliza is the modification of pronouns in recomposition. This has been used in other chatterbots but is error prone and a focus for "outing" strategies.

Eliza is basically content free and its pattern matching is too crude to be useful in a CA. Its basic trick is to ask questions which appear to be taking an interest in the user. It has no capability to sustain a conversation, but if it stalls it can retrieve an earlier user utterance and generate a new response from it.

Eliza is quite convincing because it plays a specific role with very specific demands, that of a Rogerian Therapist (Rogers, 1951). The important attributes of a Rogerian Therapist are: Congruence (genuineness, honesty with the client), Empathy (the ability to feel what the client feels) and Respect (acceptance, unconditional positive regard towards the client). Basically by reflecting back questions about past utterances Eliza appears to be taking a non-judgemental interest in the user whilst complying with the user-centred philosophy, that the client works out his or her own problems. An example of the Eliza version of pattern matching – decomposition / recomposition, is given below. Suppose the user types:

It seems that you hate me.

This would match the decomposition rule

(0 YOU 0 ME)

where 0 represents a wildcard which stores the matching substring. It would be decomposed into the structure:

1	2	3	4
It seems that	you	hate	me

And this could be processed by the recomposition rule:

(WHAT MAKES YOU THINK I 3 YOU).

Leading to the Eliza response:

What makes you think I hate you?

So in fact there is no analysis, beyond a simple surface lexical analysis and no "understanding" at all of what the conversation is about. Yet all of the set dressing makes

(a good implementation of) Eliza one of the most convincing chatterbots. Where modern implementations fail it is because too little attention has been paid to constructing a good decomposition / recomposition rule base.

1.2.1.2 Racter

Racter dates from the 1980s. It appeared as "The Policeman's Beard Is Half Constructed" (Racter, 1984) described as “the first book ever written by a computer” and as a published computer program with a compiler, Inrac, for generating conversational templates. The book, although impressive, was later suspected of being contrived with complex unpublished templates. Racter is quite convincing in having a coherent if eccentric stream of consciousness because of a heavy and repeated use of a large set of variables whose values are extracted from user utterances. It then uses a pattern matching technique to select appropriate responses to user utterances and assign values extracted from the dialogue to variables. The templates are quirky or bizarre and often contain phrases like “I was fantasising” to make the eccentricity acceptable. Racter also has methods to maintain the consistency of number and tense of verbs etc.

Racter can provide provoking and interesting responses to input user utterances. However, it does not seem to provide any coherent conversational structure and any failings in user pattern matching are masked by the overall surreal tone of the conversation.

Bringsjord (Bringsjord et al., 2001) argued that a Lovelace test for creativity is superior to the Turing Test and applied it to a storytelling machine. If anything, Racter serves as a warning to be wary of such template-based story tellers as evidence of machine intelligence.

Despite its initial impressiveness much of the knowledge of how Racter works has been lost. The book supposedly authored by Racter contains entertaining and poetic flights of fancy but not technical information. Some postings on comp.ai.nlp indicate the following pattern matching rules were supported:

Command	Description
??	Get an utterance from the user
?pattern	Search the utterance for a particular pattern
?pattern+	Search the utterance for a particular pattern, starting from the current point and moving forwards
?pattern-	Search the utterance for a particular pattern, starting from the current point and moving backwards
?pattern+[integer]	Use the integer to inspect a particular point in the utterance and compare the pattern to the word at that location
?&pattern	Match any word in the utterance ending in the pattern
?pattern&	Match any word in the utterance beginning with the pattern

/iffound
\ifnotfound For implementing if-then-else decision structures

Although the information is patchy it seems plausible to infer that the token “pattern” refers to single words or fragments. Certainly, performing the kind of pattern matching that is available in today's chatterbots would require stringing together multiple commands in a fairly complex sequence. On the plus side, the if-then-else structures could provide a surprising degree of sophistication.

1.2.1.3 Sylvie

Sylvie is a specific instance of a "Verbot" created by Virtual Personalities Inc (Plantec, 1998). This technology introduced the pattern matching rule-based system, which is familiar to most members of the chatbot community today.

Pattern matching systems consist of rules each of which contains a number of patterns and a response. A user utterance is compared with the patterns in a rule and a numerical activation is calculated (the better the match, the higher the activation). When all of the rules have been processed, the rule with the highest activation fires and its response is used to reply to the user. The patterns can contain wildcards, which allow a wider range of matches, but affect the activation produced.

The rules will often be divided into contexts to make them more manageable. It is important to realise that every time the user types an utterance, every pattern in every rule in the current context must be tested and that each test may require multiple passes through the utterance and the pattern. If a user utterance fires a rule that switches to a new context, then the whole process may have to be repeated. Therefore pattern matching is still a potentially computationally intensive process and depends on skilled context design by the people who write the pattern matching scripts as well as the software engineers who create the pattern matching engines.

The Verbot pattern matching technique was published freely during the 1990s in “The Zen of Scripting Verbots” (Plantec, 1998). The following example is adapted from Plantec. Consider a pair of activation-based rules:

<what-work>
a:0.5
p:60 What *your*job*
p:60 How *earn* living*
r:I’m a full time Verbot
+:<explain>

<explain>
a:0.5
p:60 What *you* mean*
p:60 *Eh*
p:60 *explain*
r:I am a computer program that chats with you.

Suppose the first user utterance is

“What is your job?”

Sylvie will begin by comparing it with the first pattern in the first rule:

What *your* job*

In brief the “What” sections of both strings match, but the following “is” and “your” do not. However the wildcard * is able to absorb the “is” and matching continues for the “your” substring. Wildcards are allowed to match nothing so the terminal * on “your” is ignored and the substring “job” matches. Finally the “?” symbol on the end of “job” in the utterance is absorbed by the terminal * in the pattern. This is a match which generates a positive numerical score.

The second pattern in the first rule is then tested and it fails at the first post, because the “H” in How fails to match the “W” in what and there are no leading wildcards to accommodate the difference.

When the second rule is processed the first pattern will begin by matching but will fail at the point where "job" is compared to "mean." The process continues for the remaining patterns.

The outcome is that the first rule fires (wins) and Sylvie will reply

"I'm a full time Verbot."

At this point the user will make another utterance. If there were no matches with contextually meaningful rules, Sylvie herself could fall back on a general rule that would reply with something like:

"What do you mean?"
"Sorry?”

or

"Could you explain that for me please?"

In a large base of rules the corresponding patterns (or variants on them) could occur many times so a promotion mechanism is used to ensure that the correct rule fires.

The entry +:<explain> at the end of the first rule temporarily boosts the activation for the rule <explain> for the next few utterances that the user makes. This means that a pattern in <explain> will always win over exactly the same pattern in another rule, until the boost has worn off.

Other features of pattern matching include the large range of tuneable parameters for example:

a:0.5 sets a base activation value for the rule, which is principally used to allow one of the patterns to fire when matched.

p:60 sets an activation strength for an individual pattern; this is principally used to prioritise instances of a pattern when it appears several times in different rules.

1.2.1.4 Alice

Alice is a chatterbot, which uses Alice's equivalent of a rule, known as a category. Alice categories are implemented in Artificial Intelligence Markup Language (AIML) described in a series of references and tutorials available through the AliceBot foundation web site (Wallace, 2003). A category is described as consisting of an input question known as a pattern, an output answer and an optional context. Pattern matching is similar to other chatterbots to the extent that the patterns are made up from strings (which can include spaces) and wildcards. In its simplest form the addition of a single response will complete the category. The optional context supports features such as the keyword "that" which is designed to allow Alice to remember her previous response. This allows the simulation of some short-term memory in the conversation. An AIML bot is superficially simple when compared with a Verbot. However, at the core of AIML is a simple but powerful technique - Symbolic Reduction which allows the user utterance to be successively reduced to simpler forms by re-entering the chatterbot recursively. Symbolic Reduction is intended to support the following functions:

1. Reduce complex grammatical forms to simple ones;
2. Split an input into two or more parts and combine the responses to each;
3. Match different ways of saying the same thing to the same reply;
4. Spelling or grammar corrections;
5. Detecting keywords anywhere in the input;
6. Implementing conditional branching;

   and

7. Combinations of the previous six functions.

Functions 5 and 3 are really an illustration of the limitations of basic Alice pattern matching as these come automatically with other pattern matching systems. Functions 1 and 4 are also catered for to varying degrees in other pattern matching systems. A very powerful use of symbolic reduction, which is not explicitly recognised in the documentation, is the ability to delete distracting or confusing semantic content from the utterance, revealing the parts which are actually important for pattern matching in a sensible conversation.

An example of a category using symbolic reduction, taken from Wallace (Wallace, 2003), is shown below:

```
<category>
<pattern> YOUR A * </pattern>
<template>I think you mean "you're" or "you are" not "your".
<srai>YOU ARE A <star/></srai>
</template>
</category>
```

Suppose the user utterance is

"Your a wonderful human being."

When Alice processes it with this category, the substring "you're a" matches and the substring "wonderful human being." is assigned to the variable star.

The response from the category:

I think you mean "you're" or "you are" not "your".

is displayed then the srai tags build a string and resubmit it to Alice. The string is "YOU ARE A wonderful human being." (Note that pattern matching chatterbots are not case sensitive to their inputs.)

Despite its appeal there are two problems with Symbolic Reduction. Firstly the recursion can generate a high computational load. Secondly writing a category using srai is complex and difficult to debug.

1.2.1.5 Summary of Pattern Matching Chatterbots

The previous sections give a flavour of the pattern matching approach. Of the 4 examples, Verbot (Sylvie) and Alice have persistent user communities writing Chatterbots which are usually hosted by a subscription web service. There have been no practical breakthroughs from these technologies and the clear reasons are the complexities of scripting serious

applications, difficulties of maintaining a deployed agent and the lack of support for extended goal-oriented dialogue.

1.2.2 NLP based systems

Theoretically, the main competitor to chatterbots would appear to be a Natural Language Processing (NLP) which has had huge resources expended on it over decades. Consequently, one of the strengths of NLP is that it has a substantial theoretical base. Also, particular aspects of it have been the subject of tightly focused research. However, NLP theories do not necessarily explain the workings of the human mind: for example, NLP parsers may not model the way in which humans parse sentences (Guest, 2008). This is in contrast to chatterbots which have achieved a degree of success by mimicking the human brain's abilities in feature extraction, pattern matching and forming associations.
Also, the very focusing on highly specific research topics militates against the more generalised research strategies need to build CAs. Applying NLP to dialogue has its own fundamental challenges: the usages and senses of English words are not easily identified, even a single well-formed sentence can be difficult to interpret and in an utterance the crucial information may be in one of several sentences.
NLP research also seems to have barely reached the point of considering issues of scalability to large numbers of users. This is of great importance in real-time systems deployed in the real world. This can be illustrated by two examples:

- a medium-sized airport in the UK can receive between 10,000 and 15,000 e-mail messages in a typical day (source: meeting with Manchester Airport)
- the UK national flu service received 9.3 million hits per hour on the first day of operation (resulting in it crashing even though this was a simple menu-based system) (BBC, 2009).

In fact, NLP is largely concerned with document retrieval, information extraction, and text categorisation (Jackson and Moulinier, 2007) and achievement in the field of dialogue is quite restricted.
The most promising NLP-based approach (used within a CA) currently being investigated, at the University of Cambridge, uses phrasal grammar rules to extract the dialogue act type and a list of attribute/value pairs from each utterance and a template-based response generator (Young et al., 2010, Lefevre et al., 2009). However, this approach has only been evaluated in the laboratory, with a simple domain, Towninfo, which recommends restaurants, museums and similar destinations of interest to tourists. A sample of dialogue

in Appendix 1.1 contains two examples that illustrate the relatively simple problems addressed by NLP-based systems.

Traditionally NLP systems aimed at dialogue have tended to focus on a particular area of interest of the researchers, ignoring other areas which may prove problematic in a real-world implementation. For example, Zdravkova (Zdravkova, 2000) outlined a system using parsing, keyword extraction and a structured lexicon; however there was a lack of substance when it came to reasoning about the meaning of user utterances and the production of relevant responses. Modularisation (or compartmentalisation) of NLP based systems leads to these problems being lumped together in a Natural Language Understanding (NLU) process (DeVault et al., 2009, Schmitt et al., 2009, Griol et al., 2009).

The SigDial conference series should be a good bellwether for NLP in CAs. In fact, the most recently published proceedings at the time of writing, SigDial 2009, contains just 6 of 53 papers which are on the topic of CAs (SigDial, 2009). Also, these papers are quite narrowly focused. Five are concerned with monitoring the human's engagement (Bohus and Horvitz, 2009b, Bohus and Horvitz, 2009a), interaction control (DeVault et al., 2009, Skantze and Gustafson, 2009) or determining if a party is being addressed (op den Akker and op den Akker, 2009).

A further paper investigates the effect of including CAs in multiparty dialogues as stimuli (thought-evoking). However the nature of these dialogues is extremely limited, in the form of a trivia quiz (Dohsaka et al., 2009).

SigDial 2009 contained some papers from the associated field of Dialogue Management. Although Dialogue Management is superficially identical to the field of CAs (concerned with systems designed to hold a coherent conversation with a user), individual DM researchers tend to focus on small fragmentary topics. The 2009 papers were on peripheral topics such as integration of information from pairs of sentences (Zukerman et al., 2009) and classification of a dialogue as belonging (or not belonging) to a specific type known as a "listening dialogue" (Meguro et al., 2009).

The dominant approach to NLU is the frame-based system (Minker et al., 1996, Farquhar et al., 1997, Sagae et al., 2009). These are effective for simple applications such as making bookings for journeys or theatre seats. A related approach is the use of state-based systems, popular in healthcare (Bickmore and Giorgino, 2006). These undergo state transitions triggered by the content of user utterances. In simple systems tight constraints are placed on the utterances that the users can produce. This can be done with forced choice questions (e.g. yes or no answers) or the detection of a very restricted set of highly salient

speech fragments. More flexible dialogue is possible, using more powerful grammars and probabilistic/empirical techniques, but is not trusted when high accuracy of understanding of the user intent is required (Bickmore and Giorgino, 2006).

Another recently reported system uses frame-based NLU in a training scenario, however this is described as a prototype implementation and the motivation of the researchers was to investigate overlap behaviours and the potential for a system to interrupt the user (DeVault et al., 2009).

1.2.3 Goal-Oriented Conversational Agents

The important distinction between a goal-oriented CA and an applied chatterbot is that the CA has a deep strategic purpose in holding the conversation and the mechanisms that enable it to focus the conversation on achieving a goal. The original design of chatterbots made prolonging the conversation an end in itself and thus they are easily de-railed by human users when used for practical applications.

A CA is specifically designed to interact with a human, using natural language dialogue, to achieve a specific goal - typically a business goal such as providing information, advice or selling. It plays the role of an empowered human in a productive application or task. Thus the CA may spend more time leading the conversation and asking questions than the human.

In its drive to fulfil the goal it presents with "Intentionality", that is it displays beliefs, desires and intentions concerning objects, events and states of affairs in the real world (Searle, 1999). It is not required to present as having a “mind”, but it does have to fulfil the Cartesian requirement of acting in accordance with reason (Erion, 2001).

In general terms the human approaches the conversation with a problem or need. In current implementations a pattern matching dialogue front end is combined with an expert system, which contains a model of the problem domain that is expressed in terms of a set of attributes. Through the process of dialogue, appropriate attributes are captured to model the particular problem experienced by the user and identify the appropriate solution.

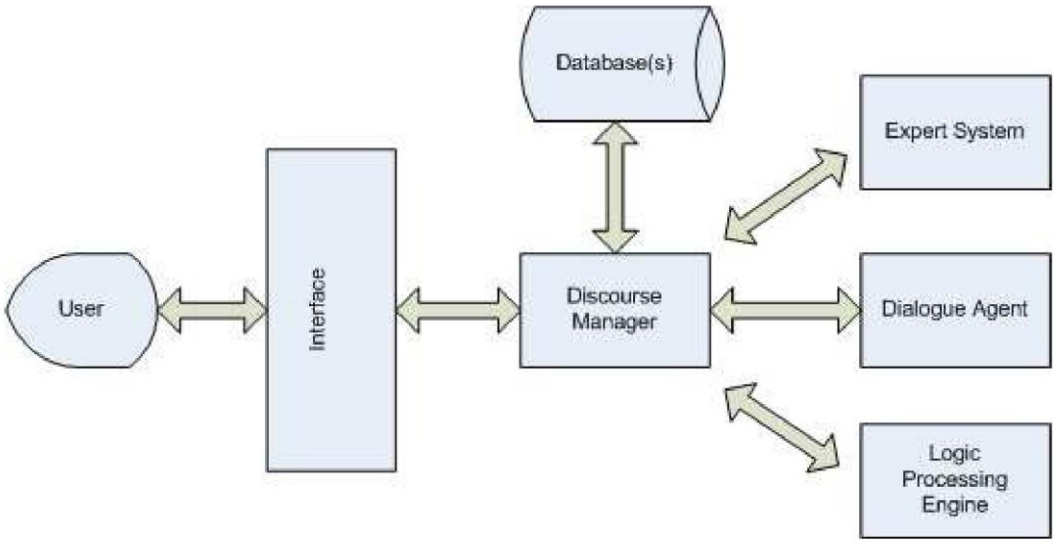
Although the CA is a mixed-initiative system (either the human or the agent may take control of the conversation), due to the goal-oriented nature of the agent it will take the initiative in the first instance and will always recover it after a brief period in which the human takes control (for example to ask for a clarification of something said by the agent).

The idea of a computer taking the role of a human in conversation was first proposed by Alan Turing (Turing, 1950), as a test of machine thought. Contemporary CAs follow Turing, in a purely behaviourist sense, in that they aspire to provide the user with the kind

of advice or services that would come from a knowledgeable or experienced human, without actually seeking to "pass" as human.

The work of the centre for Conversational Agents at MMU involves the development of a new generation of Conversational Agents (CAs) with improved techniques for dialogue management. These CAs are intended to take on challenging real-world applications in which the human user may present adversarial, disruptive or deceptive behaviour at times during the conversation (Crockett et al., 2009). Figure 1.1 shows the generic architecture for such a CA.

Figure 1.1 Conversational Agent architecture.



The architecture is best described by considering a dialogue transaction between the agent and a user.

1. The first step in a dialogue transaction is for the expert system to identify the attribute (or next attribute) whose value is to be captured.
2. The expert system passes this requirement to the discourse manager.
3. The discourse manager invokes the dialogue agent which produces a suitable utterance for the agent and returns this to the discourse manager.
4. The discourse manager passes the agent's utterance to the web interface which sends it to the user.
5. The user replies to the agent and the web interface passes the user utterance to the discourse manager.
6. The discourse manager invokes the dialogue agent to determine if it contains the required attribute.

7. If the attribute has not been captured, go to step 3.
8. When the attribute has been captured, the discourse agent passes the attribute to the expert system, which updates its model.
9. If further attributes are required by the expert system go to step 1.
10. The solution to the problem is communicated to the user.
11. At this stage the user may end the conversation or continue by asking clarification questions.

CAs have a wide range of potential applications, but they are particularly useful where the application requires sophisticated interaction with non-technical users.
Academic studies which have been conducted on CAs include health care dialogue systems (Bickmore and Giorgino, 2006), real estate sales (Cassell et al., 2000), phone call routing (Gorin et al., 1997) and intelligent tutoring (Graesser et al., 2005). CAs will be increasingly important in the future as these and similar applications are delivered remotely via the internet.
One of the most important applications for CAs is online customer self-service, providing the user with the kind of services that would come from a knowledgeable or experienced human. In 2005 there were at least 10 major companies operating in this area, including IBM and strategic partners of Microsoft (McGeary et al., 2005). At least 28 patents have been registered concerning Conversational Agents and closely related technologies. With so much investment in R&D, where are the tangible results? Commercial CAs remain basic question answering systems, incapable of genuine mixed-initiative or extended dialogue. It is now recognised that there are genuine obstacles to the transfer of CAs from the research environment to the real world (Bickmore and Giorgino, 2006).
The common feature between CAs and chatterbots is the prevalence of the technique of pattern matching. However, the important distinction between a CA and a chatterbot is that the CA has some deeper strategic purpose in holding the conversation – it is goal-oriented, whereas for the Chatterbot prolonging the conversation is an end in itself. Thus a CA will engage in extended dialogue, during the course of which it will appear to have mental states that are directed at states of affairs in the world e.g. beliefs, desires, hopes, perception, intention etc. In philosophical terms it may appear to have intentionality (Searle, 1999) - but only because this is required to achieve the goal.
This thesis focuses on Conversational Agents which serve some practical purpose and which interact with the human user through natural language dialogue (Crockett et al., 2009). The application could be giving sophisticated advice on topics such as bullying and

harassment, managing business process rules such as in the recruitment and induction of a new employee or advising on and selling a sophisticated product such as a mortgage.

1.2.4 Fundamental problems of current CA approaches

Each of the 3 approaches Chatterbots, NLP and CAs (including the Goal-Oriented versions) has fundamental problems which prevent large-scale commercial development in real-world applications:

- Chatterbots
    - Chatterbots have no practical application *per se*, because they do not achieve goals.
    - More importantly, they do not have the necessary features that would enable goal-oriented behaviour.
    - Chatterbots are supported by enthusiastic (often amateur) communities (Aimless, Wallace, 2009, Conversive) but do not draw from a substantial research base.
    - A typical chatterbot based system which is superficially appealing, but ultimately adds nothing to an FAQ list is Susan, installed on the Kegel Harley Davidson sales site (Kegel).
- NLP-based systems
    - NLP-based systems use chains of processes on the incoming text such as word sense disambiguation, parsing, stemming etc. each of which may have a certain error rate. For example, in two recent publications, parsers classed as performing well had classification accuracy in the range 86.34% - 91.13% (Ravi et al., 2008) and 92% - 93.8% (Sagae et al., 2007). Also, recent research conducted at ATT and labs concluded that "performance of all the methods for dialog act prediction is too low to be useful at the moment" (Bangalore and Stent, 2009).
    - Most parsers rely on incoming sentences being well-formed: it has been observed that "Real chat utterances are mostly unparsable. They are concerned with associative exchange of mental images. They respond to contextual relevance rather than to logical or linguistic links." (Michie, 2001).
    - NLP-based systems face a serious problem of scalability to realistic numbers of users in web-based applications, because they rely on chains of processes which can be computationally intensive.

- The issue of real-time performance of NLP-based systems remains to be tested. Even the most promising systems (Young et al., 2010, Lefevre et al., 2009) have yet to be exposed to substantial numbers of concurrent real-word users. However, information from the preparation of components such as managing training corpora (Hughes et al., 2004) or the actual training of classifiers such as Support Vector Machines (Salomon et al.) suggests that NLP processes are computationally intensive.
- Even though NLP is a large field of research, Dialogue Management is a minor area and within DM there is little enthusiasm for constructing complete CAs, as earlier noted.

- Conversational Agents
  - The production of a CA requires a substantial amount of highly skilled effort (to write pattern matching rules) in what is currently more of a craft than engineering discipline.
  - CAs are extremely difficult to maintain and this task becomes prohibitively expensive.
  - CAs are superior to established NLP approaches in real-time scalability.

1.2.5 The proposed solution – Short Text Semantic Similarity

This thesis proposes the development of a new generation of CAs based on the technique of Short Text Semantic Similarity (STSS) measurement.

This technique will overcome most of the problems arising from pattern matching which have prevented the wider-scale adoption of CAs.

1.2.5.1 Short texts – working definition

Short Texts (STs) are taken to be the length of typical user utterances in dialogue.

A typical Short Text (ST) in human dialogue would be in the range of 10-20 words (O'Shea, 2010). This definition is a matter of convenience rather than a limit on the technique and although some algorithms for processing STs (such as Decision Trees) will require an upper size limit, it can be arbitrarily large. However, common sense indicates that if a "short text" extends beyond a certain length it is no longer short and has strayed into the realm of document processing. Effectively an ST is a sentence, bearing in mind that user utterances do not always conform to the grammatical rules of sentences.

Chapters 2 and 3 of this thesis take a basic utterance to be a text string of up to 20 words in length. In chapter 4 the upper length limit was extended as far as 50 words to accommodate sentences with complex clausal structures.

1.2.5.2 Contribution of STSS to future CAs

The potential for STSS algorithms to improve CAs arises from their replacement of the pattern matching component. Suppose an STSS algorithm produces a numerical measure of semantic similarity, this could be used to make judgments such as:

- a pair of STs is identical in meaning
- a pair of STs is completely unrelated in meaning
- one pair of STs is more similar in meaning than another.

Consequently an incoming user utterance could be compared with a number of prototype statements from the domain and an appropriate action and response chosen based on the value of the best match.

Consider the following patterns, taken from a rule in a student debt advisor system:

p:15 *can*not *afford *pay*	(the p:15 parameter is for pattern
p:15 *can*not *afford *full amount*	strength, which can be varied to give
p:15 *<problem>* pay*	certain patterns higher priority, the * is
P:15 . . . . . (many more)	a variable-length wildcard)

These patterns will match utterances such as:

I cannot afford to pay you anything this term

I can't afford the full amount but I could manage to pay a third

There is a difficulty in paying because I was mugged

(amongst many others).

N.B. The second utterance matches because a simple pre-processor expands "can't" to "can not" in this architecture. The final utterance matches because <problem> is a macro which expands to include synonyms of problem such as difficulty.

It is clear that even with wildcards for generalisation, many patterns will be needed for good coverage of the overall conversational space. Also there will be a need for skilled scripters who can anticipate user utterances, generate permutations of the utterances, reduce these permutations through generalisation to patterns (use of wild cards) and, very importantly, anticipate interactions between rules.

The alternative offered by STSS is to build the rules from a set of prototype or archetype STs. Suppose, instead of patterns, we had rules containing the following STs:

I can not afford to pay.	(ps1)
My money has not come from the Student Loan Company.	(ps2)

The user utterance

I cannot afford to pay you anything this term. (u)

would be compared with all of the prototype statements using the STSS algorithms and the highest similarity match would win, as expressed in equation 1.1:

$$sim(ps1,u) > sim(ps2,u) \quad (1.1)$$

The rule containing ps1 would win and the action specified for the rule would be taken (an attribute set, response to user generated etc.)

This potential should not be underestimated. The invention of pattern matching was a paradigm shift: CAs could be created without the need for the 3rd generation procedural/imperative programming techniques in use at the time. Pattern matching, however, demanded a new mix of language and technical skills. STSS offers another paradigm shift in terms of a further reduction in the computer science skills needed in producing CAs, with a consequent reduction in development and maintenance costs. Thus with STSS we may finally see an enduring presence of CAs in real-world activities.

1.3 Related research in the field of similarity

At the onset of this work, similarity in general and semantic similarity in particular were somewhat unknown and intangible attributes. The question that had to be faced was "is similarity just some vague qualitative concept with no real scientific significance?" To answer the question a broad survey of the literature, taking in as many fields as possible, was conducted. This revealed a generalised abstract theory of similarity (Tversky, 1977), tying in with well-respected principles of measurement theory, many uses as both a dependent and independent variable in the fields of Cognitive Science, Neuropsychology and Neuroscience, and many practical applications including visual similarity and web applications. Citations for these are provided in the following sections. But perhaps the best place to start is by formalising the definition of similarity.

1.3.1 Definitions of Similarity and Semantic Similarity

We may be happy to use the words *similar* and *similarity* in real life without consciously analysing their meanings; when it comes to formalising these terms things may be a little more difficult. The following dictionary definitions are of some help:

According to the Collins Cobuild Dictionary (Sinclair, 2001):

> "If one thing is similar to another, or if two things are similar, they have features that are the same."

The definition of similarities adds a little more:

> "Similarities are features that things have which makes them similar to each other."

The Oxford dictionary (Little et al., 1983) contains several definitions of similar:

> "1. Of the same substance or structure throughout; homogeneous"
>
> "2. Having a marked resemblance or likeness; of a like nature or kind"
>
> "3. Geom. Applied to figures which may become Congruous by adjusting their linear dimensions without changing their angles"

(this is followed by examples in terms of similar triangles) and the Oxford dictionary defines similarity as

> "1. The state or fact of being similar; likeness, resemblance."
>
> "2. Points of resemblance"

So the first conclusions that might be drawn from the Cobuild definitions, Oxford definition 3 of similar and the Oxford definition of similarity are:

- Similarity arises because two different entities contain identifiable features or properties, some of which are the same
- The more of the features that are the same, the more similar the two entities are.

A less obvious conclusion from reading Oxford definitions 1 and 2 might be:

- Examining a variable attribute shared by two entities, the closer the value of the attribute the more similar the two entities are.

For example lead, gold and balsa wood all have the attribute density, but lead and gold are similar because they both have a relatively high density.

1.3.1.1 Definition of Semantic Similarity

Semantic Similarity is held to be a widely understood concept. Miller and Charles (Miller and Charles, 1991) in a word-based study wrote:

". . . subjects accept instructions to judge similarity of meaning as if they understood immediately what is being requested, then make their judgments rapidly with no apparent difficulty."

Experiments described in chapters 2 and 3 required a working definition of sentence similarity for the participants. After pilot work, this was provided by adding the question:

"How close do these two sentences come to meaning the same thing?"

1.3.2 Prior work on Similarity

This section begins by discussing what is probably the most important work on the theory of similarity, "Features of Similarity" by Amos Tversky (Tversky, 1977). Tversky is described in his obituary on the Stanford University website as "a cognitive psychologist who was a dominant figure in decision research and a leading psychological theorist." The remainder of this section reviews briefly research on similarity in a number of disciplines including Psychology, Cognitive Science, Neuropsychology and Neuroscience.

1.3.2.1 Fundamental work by Tversky

Tversky's paper (Tversky, 1977) is fundamentally important as it set out to unify the existing work on set-theoretical models of similarity into a single model. The dominant models of similarity at the time were "geometric", measuring distance rather than similarity, but always on the assumption that distance could be converted to (or negatively correlated with) similarity.

Regardless of their acceptance, Tversky criticised these models on the grounds they failed to comply with 3 fundamental axioms required for distance measures:

- Minimality
- Symmetry
- The Triangle Inequality.

Although somewhat oblique, the analysis showed that human behaviour did not support the application of the axioms.

Minimality

According to minimality the distance between any pair of identical objects should always be 0. However, Tversky observed that in certain experiments 'the probability of judging two identical stimuli as "same" rather than "different" is not constant for all stimuli.' However, the experiments mentioned were not well-reported in terms of objectives, experimental factors etc. Furthermore identification failure is known to and used by

semantic similarity researchers (Miller and Charles, 1991). Specifically, it has been used in the form of discriminability rating to produce the *d'* metric. The calculation of *d'*, accommodates the problem of higher frequency misidentification in producing a meaningful distance measure.

Symmetry

Tversky performed an experiment that showed that, under certain circumstances, human judgement of similarity was asymmetric. The standard example is the pair of judgements "North Korea is like Red China" vs. "Red China is like North Korea" (appropriate wording for the political situation at the time). The proposal was that most people find the first statement more acceptable than the second, because "Red China" is the prototype and "North Korea" is the variant. It is quite unnatural to switch the roles (as in the second version) and compare the prototype to the variant. Consequently, he proposed that in comparisons directionality is important: the variant is more similar to the prototype than vice versa. Statistically significant evidence was found to support this hypothesis in an experiment using 69 participants. This observation is relevant to the work in this thesis because the order of presentation of items in a pair presented to participants for judgement may be a confounding factor in experiments.

The Triangle Inequality

The triangle inequality problem was illustrated with the example quoted (Tversky, 1977) from William James (again an example appropriate at the time). Suppose the three entities were Jamaica, Cuba and Russia. "Jamaica is similar to Cuba (because of geographical proximity); Cuba is similar to Russia (because of political affinity); but Jamaica and Russia are not similar at all." The problem with this example is that rather than point out a fundamental flaw in distance-based measures of similarity, it may simply be pointing up the problem of immaturity in the measurement scale types when compared to those used in the physical sciences (well established measures using ratio or even interval scales). Studies of word and text similarity tend to assume ratio scale properties, with an absolute zero point (e.g. "not similar at all"(Resnik and Diab, 2000)) and an imposed upper limit. However, there has been little interest in developing an equal-interval scale, apart from some work on developing semantic anchors (Charles, 2000). At the present state of the art, we are performing the fundamental work that may one day lead to the development of such mature scales (as alchemists moved from the concepts of "hot" and "cold" to the Fahrenheit, Celsius and finally Kelvin scales).

Having criticised the prior art, Tversky's contribution was a new feature-based model of similarity – the contrast model. Equation 1.2 describes this model:

$$s(a,b) = F(A \cap B, A - B, B - A) \qquad (1.2)$$

i.e. the similarity between a and b is a function of the common features: $A \cap B$, the features in a but not in b: $A - B$ and the features in b but not in a: $B - A$.

Despite the theoretical interest of Tversky's paper, there are no semantic similarity measures based on the contrast model. Although Tversky discussed the feature-based approach and explained human perception in its terms, there is apparently no practical way to apply it to automated measures of semantic similarity. Some small-scale work has been performed using common lexical features (Jiang and Conrath, 1997), but there has been no attempt to use both the common and distinctive features.

The crux of the matter is identifying the possible features upon which a judgement would be made and how to combine (weight) the information from the salient features to produce a numerical measure.

1.3.2.2 Similarity in Cognitive Science, Neuropsychology and Neuroscience

Similarity is also ubiquitous in fields such as cognitive science, neuropsychology and neuroscience, as both an experimental control (Chwilla et al., 2007, Vigliocco et al., 2008) and an independent variable (Ballardini et al., 2007, Klein and Murphy, 2002, Jones et al., 2006).

The representation of elements of language in memory is a particularly important field. Studies include the representation of polysemous word senses (Klein and Murphy, 2001), speech production and semantic priming (Jones et al., 2006).

In these uses, it is taken for granted that similarity is a genuine attribute of objects in the world, that it can (at least to some degree) be measured or assessed and that readers of research publications have a common understanding with the authors of what similarity actually is.

Semantic priming

Semantic priming is a key investigative technique. It occurs when two words (prime and target) are presented successively to a human and the target is processed (comprehended) more efficiently because of the earlier exposure to the prime (Dien, 2008). This is believed to occur because the prime contains some of the mental code required to respond to the target (Jones et al., 2006). Priming therefore, is the result of a relationship between the two words (or terms). Often non-words are used for some of the primes to act as experimental

controls, for example by changing one vowel or consonant from a real word (Marsh et al., 2008) or as semi-random strings obeying the phonotactic constraints (permitted combinations of phonemes) of English (Vigliocco et al., 2008).

Semantic priming has also been observed with numerical stimuli (Dehaene and Naccache, 2001). An experiment was conducted in which the participants had to decide whether a number was larger or smaller than 5. When another masked number was presented as a prime, the reaction time was reduced in proportion to its similarity (closeness in numerical distance) to the target.

Category norms

Memory and language studies often use categories that are taken from an established body of prior work. These are known as "category norms"; de Zubicaray (de Zubicaray et al., 2008), in studying negative semantic priming effects, used a set created by Cree & McRea (Cree and McRae, 2003). The original category norms were created by Battig and Montague (Battig and Montague, 1969) and theirs is the best-established set, used in many studies, for example (Marsh et al., 2008 , Dennis et al., 2007, Caramazza and Shelton, 1998). This success may be attributed to the objective (Battig, 1979) ". . .these category norms may differ from numerous other similar normative projects because of our primary concern with making them as useful as possible for other researchers." Category norms are lists of words, compiled by frequency, which are made as responses by participants to a stimulus category. In a follow up study it was reported that the Battig & Montague norms have been used in over 1600 projects published in more than 200 different journals (Van Overschelde et al., 2004). Two examples of Battig & Montague categories, with members, are:

31 An article of clothing
belt, dress, shirt, sock, tie

34 A musical instrument
drum, guitar, harp, piano, trumpet

Prima facie, the members of the categories appear to be more similar to each other than to those of other categories, which is pertinent to the work in this thesis. In 2004, Van Overschelde not only updated the membership data for the existing categories, but also added 14 new categories. This points up an interesting problem for semantic similarity:

language is constantly evolving and the stability of human responses to datasets across different groups and over time may be an issue.

Battig (Battig, 1979) emphasised the value and importance of verbal materials for the research community and the difficulty in obtaining funding to produce them. The major contribution of the work in this thesis was the production of verbal materials like those of Battig & Montague. These are two benchmark datasets for evaluating new STSS measures. It is hoped that these datasets will have the same longevity in contributing to the field as those for word semantic similarity, which are still valuable 55 years after they were created (Rubenstein and Goodenough, 1965).

Neuroimaging (fMRI) studies and Similarity

Another important tool for examining brain activity is the functional Magnetic Resonance Imaging (fMRI) technique. fMRI supports the observation of the functioning of specific structures in the brain in real time. This is achieved because active areas have a higher uptake of blood oxygen (the BOLD signal) which is detected by the fMRI scanner. fMRI studies have used similarity judgement as the experimental task (Bedny and Thompson-Schill, 2006), similarity as a negative priming factor using categorically related materials (de Zubicaray et al., 2008 , Dennis et al., 2007) and as a control in deriving materials (Kuperberg et al., 2006).

Bedney investigated the effects of imagability and grammatical class during single-word comprehension, specifically whether the distinction between nouns and verbs is a true outcome of grammatical class or is an outcome of another factor, imagability.

Imagability refers to how easily a word brings to mind a visual image (and is strongly related to concreteness).

De Zubicaray was concerned with the level at which negative priming occurs within the processing of words, whether it was semantic (because it occurs with categorically related items) or structural (because items within categories share more features with each other than they do with items from other categories).

Dennis studied the effects of aging on false memory formation. Her study used categories to generate semantically related word lists. The task involved participants recalling whether or not words were part of the list and being presented with some "lures" , semantically related words from the category which were not part of the original list.

Kuperberg's study is interesting at two levels, firstly because it involved controlling the relatedness of a sequence of 3 sentences. They were selected using heuristics such as checking that one sentence could be connected to another using the word "because." These

heuristics were used to classify the texts into 3 categories: highly related, intermediately related and unrelated. The second interesting feature was the use of Latent Semantic Analysis (LSA) to investigate the consistency of the scenarios in terms of their word similarities. Comparing all content words on a term-by-term basis, the high and intermediate were not significantly different, but the intermediate scenarios were significantly different from the unrelated scenarios.

Using fMRI allows the researcher to determine whether or not there is brain activity in response to some stimulus. More importantly it also allows investigation of whether or not effects are localised to different structures in the brain.

The negative priming effect (de Zubicaray et al., 2008) provides an interesting example of the subtlety involved in these investigations. The first form of negative priming identified was the *identity* effect – when a distracter which had to be ignored in a previous trial was presented subsequently as a target, recognition was slowed. Further research showed that recognition of items from the same category as the previously used distracter was also slowed, to a lesser extent. This was termed the *conceptual* or *semantic* negative priming effect. De Zubicaray had identified the left anterolateral temporal cortex as activated during identity negative priming in a previous study, but found no evidence for this with categorical items. An effect was observed in the left posterior-mid fusiform cortex. This was considered to support feature-based processing. The formation and testing of hypotheses in the study illustrated an emerging understanding of the structure and function of the brain. De Zubicaray was able to draw on previous evidence from computational modelling, lesion and neuroimaging studies supporting a role for the anterolateral temporal cortex (ATC) in representing abstract semantic information and a role for the posterior-mid fusiform cortex in processing structural overlap or visual competition between exemplars within categories on properties of shape and configuration. This allowed a prediction that accessing semantic representations would result in activation of the left ATC and category related representations would activate the fusiform cortex. The outcome was evidence that the fusiform cortex was engaged and that the classification was feature-based.

Similarity and Category Specific Deficits - a diagnostic connection

Other studies have specifically recruited participants with impairments as a means of investigating the functionality of different brain regions. These studies typically focus on semantic memory, the ability to recall or process specific categories of words being selectively impaired by lesions affecting a particular region. An impairment of this ability is known as a Category Specific Deficit (CSD) .Two studies (Gainotti and Silveri, 1996,

Forde et al., 1997) illustrate some important considerations in neuropsychology. Both were single-patient case studies reporting evidence of a selective impairment of the category *living beings*. Forde's patient, SRB, was severely impaired in naming *fruits & vegetables* compared with his performance on inanimate knowledge, however he differed from similar patients in having no difficulty naming *food items* (outside of the *fruits & vegetables* category). Gainotti's patient, LA, also had an impairment for *living beings*; however she was spared the impairment for the *body parts* sub-category and had severe impairments in the *food items* and *musical instruments* categories. Gainotti observed that LA had bilateral lesions in both the temporo–limbic structures and the inferior temporal cortex. Whereas Forde observed SRB had a haemorrhage in the left inferior medial region of the temporal lobe extending down to the occipital lobe coupled with a small infarct in the right thalamic region. CSD studies may also control for familiarity of stimuli as well as visual complexity, and name frequency (Gainotti and Silveri, 1996).

Categories used in CSD studies are used in the work described in Chapter 3 of this thesis, so it is fair to question how the categories were originally chosen for investigation. The seminal paper in the field of CSDs (Warrington and Shallice, 1984) states that they were 26 categories selected from the 56 verbal categories in the Battig & Montague category norms study (Battig and Montague, 1969). Closer investigation shows that whilst the bulk of the categories are directly adopted some are inspired by Battig & Montague. For example there is no explicit category "Animals", however category 8 is defined as "A four-footed animal" and the category "Tools" is a similar generalisation of 23 "A carpenter's tool."

Chapter 3 of this thesis required the decomposition of words into classes. A number of techniques were considered including direct adoption of either the WordNet ontology (Miller et al., 1990)or Battig & Montague classes. However, some STSS measures (Li et al., 2004, Li et al., 2006) (Kennedy and Szpakowitz, 2008, Feng et al., 2008) use WordNet or other ontologies in calculating word similarities which were a component of the overall ST similarity measurement.

CSD evidence provides very useful tools for decomposing words into to independent, objective classes. However, there have been serious criticisms of CSDs raised in the literature. Therefore it was important to investigate the validity of these criticisms and their effect on the use of CSDs, in the following section.

Criticism of CSDs

There are a number of papers in the literature, stemming from the 1990s, which criticise CSDs on two counts. There is a third factor, more recent structural work, which is of less significance to this study, but still needs to be acknowledged.

The sceptical positions taken about CSDs were:

1. The categories do not exist at all and are artefacts of poor experimental design
2. The categories exist, but there are reductionist explanations that do not require them to be semantic categories.
3. The categories do not have corresponding neural activation sites in imaging studies of healthy participants.

Substantial and persuasive evidence has been found since the 1990s to rebut arguments 1 and 2. The third argument is still at an early stage, but emerging evidence suggests that although the neural structures may be more complex and distributed when healthy participants are observed, this does not invalidate the basic CSD concept and categories. The evidence comes in three forms large-scale reviews of prior art, re-running past experiments with tighter controls and new experiments. A review of a representative sample of this material follows.

Methodological objections to CSDs

The issue of poor design was raised in the early 1990s (Stewart et al., 1992, Funnell and Sheridan, 1992) in response to studies published in the 1980s (Warrington and McCarthy, 1983, Warrington and Shallice, 1984). This position proposed that "nuisance variables" such as familiarity, imagability, background knowledge, visual complexity, mode of stimulus and frequency of occurrence were not controlled properly across the categories being tested and were in fact responsible for the different error rates occurring for the various experimental tasks.

Since the publication of these criticisms, further studies have been conducted which have tightly controlled these variables and in which CSDs have continued to be observed (Sartori et al., 1993, Gainotti and Silveri, 1996, Forde et al., 1997) including a recent large scale review of 79 cases studies sampled (on the basis of experimental rigour) from a larger pool (Capitani et al., 2003). Indeed there are some specific cases where it has been possible to re-test a disputed case with designs to eliminate nuisance variable and the deficits have persisted, for example (Gainotti and Silveri, 1996).

Further evidence has emerged to support the robustness of CSDs. The methodological objection is undermined by the enduring credibility of the finer-grained categories (Santos and Caramazza, 2002).

Caramazza has found evidence that impairment rates are not an artefact of frequency (Caramazza and Shelton, 1998). Damasio (Damasio et al., 2004), reported that detailed analysis of the proposed confounding factors of word length, word frequency, name agreement, image agreement, familiarity or visual complexity found no systematic differences between categories that could be the basis for deficits in naming and recognition tasks. This work, using a sample of 169 patients with unilateral brain damage and 55 normal controls found evidence to support the hypothesis that the retrieval of concrete nouns belonging to distinct conceptual categories depends upon partially segregated regions in higher-order cortices of the left temporal lobe. It also found partial evidence that conceptual knowledge pertaining to the same concrete entities also depends on partially segregated regions in the right hemisphere (Damasio et al., 2004). In a naming from descriptions study, (Gainotti and Silveri, 1996), wrote "... there is considerable agreement … that category-related defects for different conceptual categories are associated with different lesion sites …"

Finally, studies attacking the existence of CSDs have themselves been subject to methodological criticism on the grounds of using (Alzheimer's) patients with mild impairments performing insignificantly challenging cognitive tasks (Whatmough and Cherkow, 2002).

Objections based on underlying models of CSDs

The reductionist approach attacks the underlying model. Although there are various models in the literature, they generally fall into one of three classes: the Sensory/Functional (SF) model, OUCH models and the Domain-Specific model (Santos and Caramazza, 2002).

According to the SF model (Warrington and Shallice, 1984, Farah and McClelland, 1991), semantic knowledge is organised into 2 modality specific subsystems. The visual subsystem stores information about visual semantic properties of objects, e.g. horses and cats both have tails and the functional/associative stores functional and other non-sensory information, e.g. guitars and banjos both have strings that can be plucked. A fundamental plank of the theory is that there is an important relationship between loss of visual/perceptual knowledge and naming impairments for living things. A consequence of

their common visual features is that living things are visually crowded compare to nonliving things and are therefore particularly vulnerable to impairment.
Consequently the model suggests that the impairments are not due to semantic properties of the word classes.
According to the Organised Unitary Content Hypothesis (OUCH) model (Caramazza et al., 1990, Caramazza and Shelton, 1998), objects have properties (animals have particular kinds of shapes, odours, motions etc) and are clustered into categories based on shared common features rather than modality. "… objects within a category share many semantic features and the semantic features tend to be highly inter-correlated."
According to the Domain-specific model semantic knowledge is organised into a relatively small number of distinct semantic categories. So patients who appear to have a general impairment for living things would be suffering from a selection of specific categories that come under the general heading of living things. This model proposes evolutionary forming and adaptation of categories, that is category-specific neural circuits were created in order to solve complex survival problems quickly and efficiently (Capitani et al., 2003). One example is the proposal by Caramazza that there is genetically based commitment of neural tissue for the recognition of animals and plants (Caramazza, 1998, Caramazza and Shelton, 1998).

Rebuttal of the objections based on underlying models

The domain-specific model supports the view that CSDs are real categories based on semantic properties.
The OUCH model is supportive at least in part. It does not propose a purely semantic categorisation but is based on co-occurrence in the world (Capitani et al., 2003). However it does propose that any category which is sufficiently conceptually compact could be subject to impairment and that all types of knowledge about the category will be affected by damage to the category. Caramazza & Shelton describe the model as having a multidimensional space of semantic properties which is "lumpy" and damage to a lumpy region impairs a category (Caramazza and Shelton, 1998).
The SF model contradicts the proposal that categories are semantic, with its assumption that the observations are the outcome of visual crowding.
From the late 1990s onwards, evidence began to accumulate which undermined reductionism. A key feature of the approach is that reductionist models inherently predict unidirectional impairments (single dissociations).

This was a fundamental strength of the SF model in explaining the early empirical results in which the living things category was impaired whilst non-living things were spared. However, as the number of single patient and small group studies has accumulated a new phenomenon, the double dissociation, has emerged. The double dissociation describes a situation in which **either** of a complementary pair of categories can be spared at the expense of the other. For example, in the large-scale review by Capitani (Capitani et al., 2003) 76 patients exhibited a dissociation for living vs. non-living things. However, of the 76 in total, 42 were more impaired for "biological categories" as opposed to 34 who were more impaired for "artefacts." This undermines the SF hypothesis which predicts all (or virtually all) impairments must be for the living category.

Double dissociations have also been observed for the verb/noun grammatical categorisation (Silveri et al., 2003). A double dissociation for concrete vs. abstract is revealed by two studies in which concrete is impaired (Breedin et al., 1994) and in which abstract is impaired (Newton and Barry, 1997) and in each case the other category is spared. This dissociation is also supported by a study of healthy volunteers which found evidence for more widespread cortical activation associated with abstract than with concrete meanings (Pexman et al., 2007).

Further evidence from Forde undermined the modality objection: the same rates of impairment were shown for fruits and vegetables whether the stimulus was visual, touch or taste (Forde et al., 1997).

Objections arising from imaging studies

Criticism has also been made on the grounds that whilst the categories exist, there is no clear link with specific brain structures when neuroimaging healthy participants (Devlin et al., 2002). Devlin reported evidence for distributed neural systems for language without functional segregation by domain or category, but also acknowledged that seven out of nine prior PET studies in literature the left posterior middle temporal gyrus responded preferentially to tools. Devlin also reported partial evidence for categories that had not reached statistical significance (Devlin et al., 2002). Moss (Moss and Tyler, 2000) reported a case study in which severe general impairment of semantic memory resulted in a selective deficit for living things, which changed to a deficit for artefacts as the patient's health deteriorated.

On the other hand Martin (Martin, 2003) reports "Lesion data and neuroimaging findings are converging to some extent in providing localisation of these components of language processing, particularly at the single-word level."

Also, Phillips (Phillips et al., 2002) reported that the left medial anterior temporal cortex was the only region to be activated for fruit relative to tools.

This debate is ongoing, but it does not call into question the existence of the categories, only the underlying neural structures that account for them.

Consequently, it was decided that the existence of CSDs is sufficiently well-established to use them in the work in this thesis.

Similarity and Modality

Neuropsychological studies also exhibit interesting variations in the presentation of items being investigated and in the instrumentation used to detect the effects, using different sensory modes. Some studies concerned with semantic properties present items to participants as pictures (Hartfield and Conture, 2006) ranging from colour photographs (Aronoff et al., 2006) and pictures (Gainotti and Silveri, 1996) to line drawings (Coslett and Lie, 2008, Slevc and Ferreira, 2006). Other studies use auditory stimuli. A study of auditory distraction in semantic memory involved the disruption of the recall of a visually presented list of items by a variety of auditory distractions including non-words, categorically related and categorically unrelated words (Marsh et al., 2008). The non-words were constructed using Sinewave speech which omits the acoustic attributes of natural speech leaving a changing pattern of vocal resonances (formants).

Eye-tracking has been used as a measure of comprehension during reading (Lee et al., 2007), timing the fixations on individual words allows time taken to comprehend a sentence in a number of different ways (including activities such as re-reading). This study investigated the interaction of the complexity of sentences and the similarity of particular words embedded in them, in sentence comprehension.

Eye fixation has also been used in an examination of the psychological validity of high dimensional semantic space models such as LSA, on the basis that people are more likely to fix their gaze on a visual referent that has some semantic relationship with a heard word, than they are to fixate an unrelated referent (Huettig et al., 2006).

Both of the above studies used an SMI Eyelink head-mounted eye-movement tracking device.

ERP Studies and Similarity

Electro-encephalogram (EEG) readings have also been used in experiments involving semantic similarity. Particular signals known as Event Related Potentials (ERPs) which are a measure of the electrical activity generated by the brain during tasks can be detected

using an EEG machine. One particular ERP, the N400 appears regularly in studies involving semantic similarity (Chwilla et al., 2007, Jones et al., 2006) and relatedness (Dien, 2008). N400 means a negative-going spike with onset 400mS following the stimulus. Dien has reported observing a variant of N400 specifically related to asymmetrically related word pairs, N400RP (an N400 signal emanating from the right parietal lobe). Compared to unrelated word pairs, there is no effect noticeable for a related word pair, but there is for an asymmetric pair where the prime reminds one of the target but not vice-versa (e.g. stork-baby vs. baby-stork).

Khader (Khader et al., 2003) investigated word category specific N400 effects (nouns vs. verbs) in German, using semantic relatedness judgement as the experimental task to generate mental processing activity. Thierry (Thierry et al., 2003) compared semantic categorisation (natural-artefact) with grammatical gender categorisation in French using auditory stimuli and monitoring several different ERPs including N4 which has some commonality with N400.

1.3.2.3 Similarity in other fields

There are a number of other fields of research in which similarity plays an important role. In some of these fields research is leading to practical real-world applications and in others research is approaching this point. These fields are summarised with brief definitions to finish this section and more detailed consideration of practical applications takes place in the following section.

Information Retrieval and similarity

Information Retrieval (IR) has been a field of study for forty years. IR was originally concerned with the search and storage of text documents, such as scientific papers, from repositories (Spärck-Jones, 1972, Salton et al., 1975). As an academic discipline it covers the development of new algorithms and processes, their evaluation in experimental settings and the evaluation of theoretical techniques in real-world applications. IR involves returning documents that are similar to either a specific example document or a search term, in each case provided by the user.

In recent years the scope of information retrieval has expanded to include text, images, sound, and DNA matching. For the purposes of this thesis it is used with its original meaning and topics such as similarity between images (visual similarity) are treated separately.

Visual Similarity

Visual similarity attempts to reproduce the ability of the human visual system to decide that two different images look similar (or not) (Liu et al., 2007b, Hsieh and Hsu, 2008). Its main application is in the management of large collections of images (such as commercial image libraries), where there may be large numbers of images of a particular interest (e.g. views of the pyramids) and where the same image could occur several times with different colour casts, sizes or qualities (compression levels). There is a strong connection with advertising, where items such as cosmetics may require searching on skin tone. Typical commercial applications use attributes such as colour, shape, texture, luminosity, complexity, objects and regions. Parameters such as percentage similarity may be used in searching the image base.

Word Sense Disambiguation and similarity

Any particular word can have several possible meanings (this property is known as Polysemy) and humans are highly effective at determining the meaning intended by another person when such a word is used, either in conversation or in writing. The individual meanings are known as senses. Computers do not have a natural facility for performing this task, so Word Sense Disambiguation algorithms have been created to perform it (Ide and Veronis, 1998). Word sense disambiguation makes use of contextual information, the local words surrounding the word to be disambiguated, and word similarity measures form part of the disambiguation process (Karov and Edelman, 1998). A particularly interesting aspect of this field is that word sense disambiguation can itself form part of similarity measurement, with word similarity used to disambiguate individual word senses which can then be used in short text measures.

Data Cleansing, Ontology Alignment and similarity

Data Cleansing and Ontology Alignment are two different, but strongly related processes, concerned with the identification and automatic elimination of errors in real-world data sets. Data Cleansing is associated with the merging of records from two or more large databases (Maletic and Marcus, 2000). On the other hand Ontology alignment is largely concerned with interworking between separate ontologies consisting of entities, attributes and rules (Stoilos et al., 2005). Real-world collections are both large and have relatively high error rates, making these activities extremely important. A related area is the Semantic Web.

The Semantic Web and similarity

The Semantic Web is a framework for sharing data located in ontologies on the web between different users or different applications. The distinguishing characteristics of the semantic web are highly distributed data and diverse ownership. The semantic web is associated with Materials Requirements Planning and supply chains (Anjomshoaa et al., 2006). In particular it concerns abstract forms of component descriptions, order documents etc which allow easy interaction between manufacturing businesses in the supply chain. Once again, similarity has an important role to play through discovering semantically similar entities in different ontologies spread through the semantic web.

XML and similarity

XML is a set of rules for encoding documents electronically, which is strongly associated with documents used on the web. There is a significant interest in measuring the semantic similarity of XML documents (Jeong et al., 2008). This again is related to the business-to-business communication via the web described in the previous section.

1.3.3 Real-world acceptance of similarity

Whatever difficulties there may be in formulating similarity on a rigorous scientific basis, we are nonetheless happy to use the concept in high-stakes applications in the real world. For example in fingerprint matching (Joun et al., 2003) where similarity is calculated using types, angles and corresponding positions of distinctive features known as "minutiae." A second example is the FBI CODIS database (Rieck and Laskov, 2007) of over 5,000,000 offender profiles, where matching works by looking at counts of repetitive short sequences that occur at known sites (STR loci). In both these applications, data volumes are so high that only machine measures can produce results in a useful time-frame. The method reported by Joun uses Bayesian probability modelling and that by Rieck & Laskov treats the data as symbol-strings and uses string matching measures. The following sections explore the potential for practical applications in the fields described in section 1.3.2.3.

1.3.3.1 Visual similarity and its practical applications

A practical application of visual similarity is in the management of still and moving images. Large collections of images, whether still or moving, are complex to manage; collections are diverse and often poorly indexed (Zhang and Zhang, 2006). Such collections may use similarity retrieval to find the database images that are similar to the given query image (Hsieh and Hsu, 2008, Zhang and Zhang, 2006, Liu et al., 2007b). Image retrieval can exploit similarity at a number of levels. These range from basic extracted features such as the colour, texture and shape information (Rallabandi and Sett,

2008) to high level conceptual descriptions. The *semantic gap* (Liu et al., 2007b) occurs in moving from feature-based retrieval to semantic retrieval; the first step being to queries such as "find a picture of a flower" and from there to more high level queries involving emotional states etc. such as "find pictures of a joyful crowd."

The RETIN search engine architecture (Gosselin et al., 2008) uses a multi-stage approach in tackling the semantic gap. The stages comprise extraction of colour and textural information, indexing, formation of feature vectors in a Hilbert space by applying a kernel function, active learning (through user labelling of retrieved images).

Image similarity may be applied to find the boundaries of shots in a movie. Approaches include integrating low-level features (colour and texture) to compute the similarity between shots (Chen et al., 2008) or computing changes of text and background regions in videotaped lectures containing slide presentations (Wang et al., 2008). Yi also uses segmentation as part of an overall content management system (Yi et al., 2006). Segmentation is followed by collection of shots into scenes (described as having a meaningful semantic content) and finally retrieval based on similarity between the semantics of the video content and the desired semantics.

A number of movie applications use labelling and term semantic similarity.

Wang's application is interesting because a purely visual process is followed by a lexical semantic measure for matching shots to documents.

A natural language querying system for video databases (Erozel et al., 2008) considers semantic content to consist of objects, activities, and spatial properties of objects. Queries can take forms such as "Show all frames where Bill is seen at the upper-left of the screen." All objects are annotated as nouns and all actions as verbs in a database. The natural language queries are converted to formal semantic representations and used to search the database. Semantic similarity is used in two ways, ontology-based querying and query substitution. If an exact match to a query is not found, ontology-based querying uses WordNet to find the most similar concepts in the database to those in the query and returns approximate matches. Query expansion again uses WordNet, but this time to generate permutations of similar queries which can be submitted to the database. The word similarity measure used was a version of Wu and Palmer's method (Wu and Palmer, 1994).

An alternative video querying system developed at MIT (Katz et al., 2003) focuses on answering natural language questions about moving objects. The natural language component uses the START system, developed for information retrieval, making use of annotation, grammar, syntax, rules and a lexicon but not semantic similarity.

Methods of computing visual similarity include a feature-based Euclidean distance measure (Zhang and Zhang, 2006), graph measures (Hsieh and Hsu, 2008, Shokoufandeh et al., 2006), self-organising maps (Rallabandi and Sett, 2008) and histogram methods (Chen et al., 2008, Yi et al., 2006).
Methods of computing visual similarity include feature-based Euclidian distance measures (Zhang and Zhang, 2006, Iyer et al., 2005), a feature-based graph measure (Hsieh and Hsu, 2008) and feature-based self-organising maps (Rallabandi and Sett, 2008).

1.3.3.2 Practical applications of Semantic Similarity in NLP and Information Retrieval

Applications of semantic similarity in NLP tasks (Montemagni and Pirrelli, 1998) include "syntactic disambiguation (either structural or functional), word sense disambiguation, selection of appropriate translation equivalent, assessment of lexical cohesion in texts for automatic summarisation, query expansion and document indexing in Information Retrieval."

1.3.3.3 Question Answering and Information Retrieval

Question Answering (QA) and Information Retrieval (IR) are intimately related, overlapping fields. It is not much use having the mechanism to identify a question and what it is about, if there is no means of acquiring the answer. Conversely, the ability to retrieve information seems of little point unless there are reasons to retrieve it. The distinction arises because, historically, IR was concerned with the retrieval of documents, such as scientific papers, from repositories, whereas QA is concerned with more generally phrased queries (often looking for specific facts) and these days focuses on the World Wide Web as the source of information for formulating answers.
Systematic research in retrieving appropriate subsets of documents from sizable collections has taken place since the 1970s (Spärck-Jones, 1972) and has origins in research going back to the 1960s (Masterman et al., 1959). To the scientist, the most obvious application of IR is searching databases of scientific papers such as INSPEC, COMPENDEX or Web of Knowledge. IR is equally important in retrieving authors' self-archived publications from the more distributed and less structured World Wide Web, using search engines such as CiteSeer and Google Scholar. The specific form of semantic similarity, term-to-document, is fundamental to IR. A non-proprietary agent, CiteSeer, which is discussed in the literature (Bollacker et al., 1998) uses a combination of semantic distance measures in answering queries. Variant citations of the same paper are discovered using edit distance and term frequency measures. To find a set of similar (but different) documents CiteSeer

uses a TF*IDF (Term Frequency – Inverse Document Frequency (Spärck-Jones, 1972)) measure, a string distance measure (called LikeIt) and an adaptation of TF*IDF to exploit commonality between citations in the documents (called CCIDF). TF*IDF is described in section 1.4.3.2.

One particular aspect of Information Retrieval, Question Answering, is also highly relevant as an application because it has several stages in which semantic similarity measures can be used. The TREC QA (Text Retrieval Conference Question Answering) track is a good illustration of the importance of semantic similarity because in the period 1999-2007 at least 57 universities and at least 20 business entered systems for evaluation. The Universities include Imperial College, MIT and Columbia University. The businesses include Microsoft, IBM and AT&T. The US government considers text retrieval to be of high strategic importance. The TREC conference series is sponsored by NIST (National Institute of Standards and Technology) and IARPA (Intelligence Advanced Research Projects Activity), both US Government agencies. The IARPA mission statement (IARPA) contains the following:

> "The Intelligence Advanced Research Projects Activity (IARPA) invests in high-risk/high-payoff research that has the potential to provide our nation with an overwhelming intelligence advantage over future adversaries."

Additionally, the US National Security Agency has been an active participant in the TREC QA track. Some examples of intelligence-related questions used in test collections for the track (Katz et al., 2005a) are:

> "What agent type did Aum Shinrikyo use to execute an attack in Japan?"

And

> "What countries were bases of operations for Al-Fatah in 1980?"

Semantic similarity measures contribute to several of the stages involved in QA systems. These usually begin with question pre-processing to form a query. Some systems begin with query expansion, in which multiple versions of the query are creating using WordNet synsets of the terms in the original query (Quarteroni and Manandhar, 2006).

The next step involves retrieving a subset of documents from the collection that might contain the answer to the question. This involves (explicitly or implicitly) some form of query-document similarity measurement. The nature of the query means that a simple vector-space similarity measure is capable of retrieving a set of documents containing potential candidate answers from a realistically sized document collection. For example,

PRISE is a large vector space model, which treats the documents and queries as a list of words and returns a subset of documents from the collection ranked in order of their statistical similarity to the query (Dimmick, 2000). Prise is freely available from the National Institute of Standards and Technology (NIST) (NIST, Dimmick, 2000).

Having found documents which contain candidate answers to the question, it is necessary to extract the relatively small pertinent material from the rest of the document. The pertinent unit, referred to as a "snippet" may be a sentence or a short paragraph-sized passage (Katz et al., 2007). Similarity measures may be used in this extraction process, one example being a WordNet based measure used during the development of the YorkQA system (De Boni and Manandhar, 2003). Some systems work at an even finer-grained level, the "nugget", when performing answer extraction. Nuggets are pieces of knowledge at the phrase level.

Some systems use a separate ranking step incorporating similarity. An example of the variation that can occur within this basically common approach, a web-base question answering system from MIT CSAIL (Katz et al., 2005b) used the BLEU metric to identify nuggets from sources such as Wikipedia, followed by TF*IDF similarity measurement to rank candidates for relevance.

The final stage, answer formulation, also offers opportunities for similarity measures to play a role. The set of documents returned by the IR stage may contain redundancy. CSAIL overcome this with a similarity-based novelty component in the candidate ranking stage (Katz et al., 2007).

Filtering has even greater relevance to dialogue-based question answering. As TREC evolved, it introduced the task of dealing with a series of questions about a target, the final one being the "Other" question. Dang (Dang et al., 2007) interprets it as:

> "Tell me other interesting things about this target I don't know enough to ask directly".

TREC 2007 emphasised the importance of not repeating information which had been covered in answers to earlier questions in the series (Dang et al., 2007), making the task more difficult. A simple approach taken by Intellexer (Bondarionok et al., 2007) stores the sentences used to answer previous factoid and list questions, then trims them from the list used for answer extraction. Use of a semantic similarity measure could reduce the size of the list, whilst improving its coverage of paraphrased sentences.

1.3.3.4 Copyright and Plagiarism

This is another area of interest to researchers at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). Their approach to detecting copyright infringement (Uzuner et al., 2004) was shaped by US copyright law which protects the form of expression of material but not the underlying facts or ideas. Consequently, they evaluated syntactical or style features as a method of identifying infringement and compared its efficacy with content similarity. In 2004 they used decision trees trained with 23 different features from several different categories: Surface (e.g. average word length), tagged-text (e.g. distribution of unique POS tags), parsed-text (e.g. frequency of imperatives), Semantic (e.g. frequency of uncertainty markers such as "maybe"). Interestingly from the point of view of this thesis, the syntactic similarity measure showed a fair level of agreement with the content similarity measure which used TF*IDF-weighted unigrams. Combination of the two measures in general showed an improvement indicating use of different information. The approach to experimental evaluation is also interesting. Human evaluators rated pairs of news articles (from a repository in which articles had been deliberately re-written with different levels of similarity). Participants evaluated (on average) 6 pairs of documents given 15 minutes per evaluation. They were asked to rate the pairs using a scale running from not similar = 0, slightly similar = 0.2, somewhat similar = 0.4, pretty similar = 0.6, very similar = 0.8, to completely similar = 1. They also provided the featural information used to build the decision trees. The study referred to the human judgements as "ground truth." However, there is no objective measure to establish ground truth in semantic similarity so this term should be used with care.
Further work published in 2007 used multiple translations of the same literary work (e.g. Madame Bovary) as the data set. This study drew its features from sentence-initial and sentence-final phrase structures, plus 3 groups of features derived from verbs. The theoretical background to the verb feature structures was derived from Levin's alternation-based verb classification system (Levin, 1993). This study did not require human participants as it was already known which were the paraphrased passages.

1.3.3.5 Word Sense Disambiguation

Word Sense Disambiguation (WSD) is not an end in itself, but is an important component of a number of NLP processes. Disambiguation by machine has been an area of interest since the 1950s (Locke and Booth, 1955). By 1998 (Ide and Veronis, 1998) it was recognised as an essential component in natural language understanding applications such as message understanding (for example the automatic processing of e-mail messages) and human-machine communication (potential NLP-based conversational agents). Ide and

Veronis (Ide and Veronis, 1998) also identified other areas which WSD can contribute to, including Machine Translation, Information Retrieval, Content and Thematical Analysis, Grammatical Analysis (e.g. in POS tagging), Speech Processing and Text Processing (e.g. in spelling correction). A similar list was also given in (Heylighen, 2001).

Experience at the MMU Intelligent Systems Group suggests that, even when using pattern matching CAs without NLP processes, considerable effort is required in writing scripts to resolve ambiguities.

By the mid 1990s word semantic similarity algorithms were being applied to WSD. For example, a conceptual distance measure in a WordNet hierarchy used path length and depth, relating the possible sense of the current word to its context, a small moving window surrounding it in the text (Agirre and Rigau, 1995). This algorithm also takes account of conceptual density, in treating concepts in denser parts of the hierarchy as closer than those in sparser parts of the hierarchy.

Since 1998 the Senseval / Semeval conference series has been a focus of research (Mihalcea and Pedersen). In a manner much like TREC, this conference series provides test collections and a framework for the evaluation of WSD programmes. Since 2007 there has been a migration from purely WSD to a broader range of semantic analysis tasks. One such task is finding substitute words, closely associated with word sense disambiguation, as it involves finding semantically similar words that fit the context of the original word (Hassan et al., 2007). Hassan lists Lexical Acquisition, Question Answering and Text Simplification as applications in addition to some of those listed earlier for WSD. The approach uses a number of candidate ranking algorithms for semantic relatedness (including LSA) and language modelling (to ensure the substitute is a good fit) combined through a voting mechanism.

In recent years good quality parallel corpora have become available; parallel corpora contain equivalent texts in different languages. This provides a rich source of information for machine WSD algorithms because words do not fall into overlapping sets of polysemous senses in different languages, therefore the translator has already performed the work of disambiguation (Tufis, 2004). Tufis et al report that diversity and volume of data is important and that using 6 languages from 3 different groups (Romance, Slavic and Finno-Ugric) resulted in 75% accuracy in sense-clustering English words. The similarity algorithm relies on semantically aligned WordNets in the different languages. Then, for a pair of terms, the similarity is computed in terms of links between the members of parallel synsets or links to the nearest common ancestor.

Semantic similarity may be used in two ways for WSD; in a local context, between pairs of word or in a global context in lexical chains (Mihalcea and Pedersen, 2005). The local context involves calculating the similarity between the current word and the senses of words to the left and right. Mihalcea and Pedersen's example being that in the phrase "plant with flowers", the sense flora has a higher similarity to flowers than the sense of industrial machinery.

In the global context, a sequence of candidate words is selected for disambiguation. Working through the sequence each word is disambiguated based on its semantic similarity to the senses of words which have previously been disambiguated in the chain. Mihalcea and Pedersen's example in this case being that in the sentence beginning "A very long train travelling along the rails . . ."

- Train is disambiguated as public transport (as opposed to an order of things, a piece of cloth)
- After which Travelling is disambiguated as change location (as opposed to undergo transportation)
- And consequently Rail is disambiguated as a bar of steel for trains (as opposed to a barrier, a small bird, express objections or criticisms in abusive language etc.)

Lexical chains have been reported as achieving between 60% and 90% accuracy.

A philosophical and forward-looking study by Heylighen (Heylighen, 2001) investigated the potential value of large-scale associative networks, constructed through web-mining for disambiguation. Heylighen pointed out the lack of semantic primitives (grounding) from which meanings can be constructed, emphasising that the meaning of a word is derived from the whole of its associative relations with other words. This work is concerned with associations rather than similarities between words. Although it is pointed out that one way to measure similarity would be through association strength, he pointed out that the vector-space method is likely to be more reliable. Heylighen also observed problems of symmetry (mouse is more strongly associated with cheese than vice-versa) and transitivity (cat is strongly associated with mouse but following a link from cat through mouse to cheese falsely implies a strong association between cat and cheese). This is a re-iteration of two problems discussed by Tversky (Tversky, 1977). He also noted that the relative values of associations are more important than absolute values when it comes to practical linguistic tasks.

One experiment reported in the paper involved clustering the 150 most frequent nouns using the association matrix – this grouped most of them into one of 9 superclasses.

Blurring between association and semantic similarity occurs in the discussion of other work where an association matrix was used to model semantic priming, where "…the degree to which one word can prime another word in a psychological experiment is correlated with the degree of semantic similarity between word vectors." [from the association matrix]. Similarity occurs more explicitly in the process of disambiguation. This involves creating a spreading activation from the association matrix, so that words occurring in the text activate neighbours (with portions of the activation being passed on to successive neighbours of activated words). The similarity between the activation vectors for the word senses of the ambiguous word and the activation vector of the surrounding words can be calculated, using the most similar sense vector to select the actual meaning of the word. Heylighen also proposed modified association network which could be used to optimise links between pages and to create web indexes.

1.3.3.6 Data Cleansing and Ontology Alignment

Similarity measurement has a role in the identification and automatic elimination of errors in real-world data sets. Even with reasonable diligence at the data capture stage error rates of 5% have been reported (Maletic and Marcus, 2000). The application of data cleansing is particularly important in merging (Maletic and Marcus, 2000) or linking large databases (Liang et al., 2003).

Maletic and Marcus (Maletic and Marcus, 2000) used similarity measures to form clusters and consequently identify outlier records. Two approaches were used, Hamming distance based on empty / non-empty and the Euclidean distance between fields in the records.

A more thorough approach by Liang (Liang et al., 2003) emphasised domain specific similarity and split the problem into two levels: defining similarity metrics for each field in the record and defining a function for combining the individual field measures. String edit distance was described as a basic approach at the field level, but it was proposed that different functions could be developed for data such as names and addresses. An alternative n-gram (at the character level) field measure underlines the position of Liang et al that most errors will be typographical during data entry. *An n-gram is a sequence of contiguous symbols found inside a longer sequence. The form of n-gram which is most relevant to this thesis is the word n-gram, a sequence of words extracted from a short text or a document.*

The similarity measures described are highly simplistic compared to semantic similarity. It is possible however, that semantic similarity could contribute to this process. Short text

similarity, in particular, could contribute to applications where the databases contain free text fields.

Ontology alignment is a related area of work. An ontology is "a rigorous and exhaustive organisation of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations" (Princeton).

Formal Concept Analysis (FCA) is a technique for automatically constructing Ontologies from a collection of entities and their attributes. It discovers clusters of entities and the clusters of their shared attributes. Similarity between concepts has been identified as an important process for FCA (Formica, 2008). It is particularly important for ontology merging and alignment which Formica described as time consuming and error prone. This study uses the information-based semantic similarity measure from Lin (Lin, 1998). In FCA a concept has an Intent and an Extent, and sets of object and attribute components. In Formica's proposal, similarity between concepts is calculated as the weighted sum of similarities between the Intents of the two concepts and the Extents of the two concepts. Whilst this work is an interesting proposal, there is no significant evaluation, only a few worked examples are given.

Perhaps the most interesting application of STSS to databases is in allowing queries to be expressed in natural language, although this field is still in its infancy (Glass et al., 2004, Owda et al., 2007).

1.3.3.7 The Semantic Web

The Semantic Web is a framework for sharing data located in ontologies on the web between different users or different applications.

The key role of ontologies was identified by Hu et al. (Hu et al., 2008), who observed that there is a proliferation of ontologies on the web (in excess of 10,000) and that a number of ontologies may deal with overlapping or even identical domains. Interoperation is desirable to exploit the full range of information available and this requires alignment (discovering semantically similar entities in different ontologies). Again the simplest approach is string matching through edit-distance type measures. However Hu reported these have been supplemented with domain knowledge in fields such as biology and geography. Hu et al. used a string metric similarity metric proposed by (Stoilos et al., 2005) which takes account of differences as well as similarities (corresponding to Tversky's contrast model). Hu et al developed the process by partitioning the ontologies into blocks and finding high similarity alignments between blocks known as anchors. In discovering anchors, they looked for similarities between entity names using an

improvement on edit-distance metrics. A cosine measure of similarity between structures in ontologies was also used to tune the alignment measure.

A related piece of work, falling between Ontology Alignment and Information Retrieval involved constructing hypertext links within and between newspaper articles (Green, 1997). This work represented text units at the paragraph level in a document by lexical chains (sequences of semantically related words in the paragraph); the chains having density values which indicate their importance to the paragraph. A simplistic 3-level WordNet-based measure was used for relatedness between words in a chain, the levels being Extra Strong (identical word), Strong (in the same synset) and Regular (connected by a path with a maximum of 5 links).

Similarities between lexical chains were calculated using a similarity coefficient (the dice coefficient was used in the publication). HTML links were then inserted based on the similarity calculations. A vector-space measure of similarity between documents was used for generating links between units of text in different documents. This measure had a vector whose elements corresponded to WordNet synsets and whose elements contained weights calculated from the frequency of occurrence of the synset in the document. Weights were calculated using a function from Salton's SMART algorithm (Salton and Allan, 1993).

An ontology enrichment process was described by Faatz and Steinmetz (Faatz and Steinmetz, 2002). This involves building a text corpus by querying the web using terms extracted from the ontology, then extracting statistical information from the corpus to provide contextual information about the terms. The intended outcome is the automatic discovery of new concepts (from the corpus) which extend the ontology. Similarity measures play a role in selecting candidate concepts according to their similarity to concepts already in the ontology. Semantic similarity is measured between the verbal descriptors of concepts and takes the form of a distance measure between weighted, sparse vectors.

There is significant interest in a process known as "ontology mapping" in which the original ontologies remain unchanged. Su and Gulla described the challenges raised by working with the semantic web in which numerous ontologies describe the semantics of a large quantity of data distributed throughout the web (Su and Gulla, 2006). They related that there is no single shared ontology that can be adopted by all users, hence the need for interoperability between heterogeneous systems. Their method, described as "semantic enrichment", calculates the similarity between pairs of concepts or pairs of relations in two different ontologies. Enrichment involves building enriched feature vectors for the

concepts in the ontologies by incorporating information from document sets associated with each ontology (documents are associated by a document classifier stage). Similarity was calculated as the cosine between vectors, with an adjustment stage involving a path-length-based WordNet similarity measure, preceded by POS-tagging. Su and Gulla specified that this stage covered only nouns, verbs and adjectives.

Another approach to computing semantic similarity across separately maintained ontologies (without constructing a shared ontology) used a composite similarity measure (Rodriguez and Egenhofer, 2003). One interesting feature of the measure was that Rodriguez and Egenhofer adopted Tversky's model in order to accommodate their view that asymmetry is important in this application. The similarity model itself had 3 components: word matching (common and different words between synsets denoting the entity classes), feature matching (between feature sets of the entity classes) performed with string matching algorithms and semantic neighbourhood matching. This was based on the number of similar classes within the semantic neighbourhoods of the two classes being compared. The semantic neighbourhoods themselves are derived by synset or feature matching and make use of path length.

Rodriguez and Egenhofer were critical of prior work on semantic similarity stating "There are few studies that have addressed the quality of results of similarity assessment." The exceptions they noted were Jiang & Conrath, Miller & Charles and Resnik. However it should be noted that their own work contained a very small evaluation (a task taking a maximum of 10 minutes) and there would be serious questions about scaling up their small ad hoc ontology to the real world.

1.3.3.8 Other work involving the Web and Similarity

In a survey of web-based metrics (Dhyani et al., 2002), 3 different classes of web page similarity metrics were identified: content-based, link-based and usage-based. They further decomposed content-based measures into term overlap and vector-space methods. The link-based methods were described as derivatives of bibliographic measures based on shared citations or co-citation (where a third paper cites both papers being compared). Usage-based methods derive their information from web server logs.

Since the inception of extensible markup language (XML), which allows documents to be marked up and shared via the web, there has been an interest in measuring the semantic similarity of XML documents. There is particular interest in using this facility for business-to-business communication via the web, for example communicating process through a supply chain.

Semantic similarity has been used for matching XML schemata, by combining established similarity measures using neural networks (Jeong et al., 2008), string similarity, token similarity and token synonym similarity (Saleem et al., 2008). A variety of methods have been evaluated for increasing interoperability of data exchange through XML messaging, including LSA, tree edit distance, cosine measure and two kernel string similarity functions (Jeong et al., 2009). At a fine-grained level similarity has played a role in clustering of segments of XML documents (Langer et al., 2004), using the K-nearest neighbour technique. The related task of integrating web taxonomies has used a cosine measure for extracting fine-grain semantic relationships (Wu et al., 2007).

The processes of clustering and classifying across a range of web materials are of interest in their own right. Grimnes (Grimnes et al., 2008) used a graph distance measure to cluster objects from the semantic web, Roussinov (Roussinov and Zhao, 2003) used construction of a WordNet-like ontology to cluster messages generated through electronic brainstorming sessions and Sahami (Sahami, 2006) used a Cosine measure to find recommendations for web communities for users to join. Sahami's study went beyond social networking to consider record linkage and short text similarity in general. Another feature of this work was query expansion – expanding the semantic vectors with contextual material found by search engine lookup. Query expansion was also investigated by Zelikovitz (Zelikovitz and Hirsh, 2000) for short text classification. This used a classic TF*IDF cosine measure, but expanded the semantic vector using a joint set of unlabelled background knowledge. This is a recurring theme throughout activities like question answering - how to exploit the enormous base of information available through the web and other ontologies without performing large scale human classification or labeling of the data. Another technique, Approach to Determining Semantic Similarity (ADSS) (Han et al., 2006), also exploits contextual information for ranking interconnected entities returned by a web search (which the authors describe as more difficult than straightforward document ranking). This technique computes similarity using extracted features.

A verb similarity measure from Wu & Palmer has been adapted for linking (Halkidi et al., 2003) and indexing (Desmontils and Jacquin, 2002) websites. Wu & Palmer's approach calculated similarity using a path length measure in a conceptually structured ontology that made use of Levin class properties such as motion, force, contact, change-of-state and action (Levin, 1993), so clearly some adaptation is needed for their measure to cover other word classes. Both Halkidi and Desmontils use the Wu & Palmer measure with WordNet.

Two quite unusual approaches round off this section. The first is intended for use with very short texts (2-3 words) and is intended for search engine queries. This technique (Chien

and Immorlica, 2005) makes no use of the verbal content of the terms, but computes similarity by correlating temporal features (frequency functions). Example very short texts are *weather channel* and national *weather service*.

Pointwise Mutual Information and Information Retrieval (PMI-IR) was a technique originally proposed by Turney (Turney, 2001b) to use semantic similarity to find synonyms on the web. This was followed by a debate as to whether the technique could find good answers to subcognitive questions posed in a Turing Test (French, 2001, Turney, 2001a). This technique relied on the AltaVista search engine. Investigative work for this thesis found the feature in AltaVista to be either removed or no longer working reliably. Correspondence with Turney indicated that there was no alternative strategy. Recent work by Inkpen (Inkpen, 2007) suggests that a variant of PMI-IR, SOC-PMI, (Second Order Co-occurrence PMI) may be viable using smaller corpora than the web, such as the British National Corpus (BNC - which is a large corpus in its own right, but smaller than the web).

1.4 Key Concepts

This section discusses some concepts which are crucial to the work in the following chapters. In the first instance it considers whether STSS algorithms should be tested using grammatically correct sentences or with the more haphazard utterances that occur in human conversation. It continues with the distinction between similarity and relatedness, which are sometimes used interchangeably, examines the background of existing text similarity measures and summarises prior work on evaluating CAs. This leads to important questions which must be addressed to evaluate STSS measures.

1.4.1 Short texts or sentences?

The work in this thesis springs from the new STSS algorithm (now known as STASIS) developed by the MMU intelligent systems group, first published in 2004 (Li et al., 2004). Because the application was intended to be in dialogue, where user utterances are often grammatically incorrect (Michie, 2001), the algorithm was designed not to depend on receiving properly-formed sentences as input. Also, the only other algorithm at the time that could be applied to short texts, LSA, did not require sentences as input. (The LSA portal has an option for "sentence-to-sentence" comparison but does not require the two "sentences" to be valid.)

At that time it was anticipated that the NLP community would respond with alternative measures using techniques such as parsing, disambiguation etc. that would require properly-formed sentences. Therefore, for two reasons it was decided to seek valid

sentences when collecting the materials to form the Gold Standard datasets developed in this thesis.

The first reason was to facilitate fair comparison between NLP-based and non-NLP based algorithms, by factoring out the stage of sentence-repair that would be required for NLP-systems (as automatic sentence-repair algorithms may introduce their own errors). The second reason was for the benefit of participants, who could find "hedging" instructions which were asking them to produce something like a sentence, but not necessarily a correct sentence, somewhat ambiguous.

In fact, virtually all of the STSS work since 2004 has been influenced in some way by either STASIS or LSA and consequently has not required grammatically correct sentences. However, these datasets leave the door open for future developments using NLP to be compared with other algorithms. Furthermore, explanation of the operations of STASIS and LSA will show that current algorithms are not given any unfair advantage by restricting the dataset to valid sentences.

1.4.2 Similarity vs. Relatedness, Antonymy and kinds of similarity

Empirical studies suggest that semantic similarity is a little more subtle than has been assumed. Some draw a distinction between "similarity" and "relatedness" (Resnik, 1999, Vigliocco et al., 2002a). Resnik gives an example: cars and gasoline seem more closely related than cars and bicycles, but the latter pair is more similar. Although Resnik specifies semantic similarity as a special case of semantic relatedness, Charles has used relatedness to describe degrees of similarity in an empirical study (Charles, 2000).

Four forms of similarity are described by Klein and Murphy (Klein and Murphy, 2002): Taxonomic, Thematic, Goal-derived and Radial. Taxonomic similarity is the foundation of Noun similarity studies, following ISA relations through a structure such as WordNet. Cars and gasoline are a good example of Thematic similarity (related by co-occurrence or function). Goal–derived items are connected by their significance in achieving some goal and Radial items are connected through a chain of similar items, possibly through some evolutionary process. The context in which the similarity judgment is made could result in any of the forms dominating the decision.

In some studies Semantic Distance (difference) is measured. Distance can be thought of as dissimilarity - the counterpart of semantic similarity. So if a study measures distance, it is taken as having measured similarity, by applying an inversion operation (Tversky, 1977) or by looking for a negative correlation with distance instead of a positive correlation with similarity (Miller and Charles, 1991).

The concept of similarity may in itself be asymmetrical, depending on the circumstances in which items are presented. According to Tversky, "A man is like a tree" and "A tree is like a man" are interpreted as having different meanings (Tversky, 1977). Gleitman et al (Gleitman et al., 1996) claim that the structural position of the noun phrases set them as figure and ground or variant and referent, leading to the asymmetry.

Most studies use similarity measures on a scale running from 0 to a specified maximum value, typically 4. However this rating scale has no capacity to represent oppositeness (antonymy) as more different than having no similarity at all. Antonyms also generate high similarity values with co-occurrence measures (Miller and Charles, 1991).

1.4.3. Origins of Semantic Similarity measures

1.4.3.1 Euclidian Distance

A Euclidian distance measure assumes that the items (e.g. sentences) are located at two distinct points (*p, q*) in an *n*-dimensional semantic space. Each dimension of the space represents an attribute used in judging similarity. Calculating the Euclidian distance (equation 1.3) between the two points provides a measure of *semantic distance (d)*.

$$d = \sqrt{\sum_{i=1}^{n} [p_i - q_i]^2} \qquad (1.3)$$

Distance is treated as an unsigned scalar value and consequently similarity measures influenced by this model are unipolar and unsuitable for antonyms. This model may prove useful in qualitative descriptions of similarity but it is difficult to operationalise because we have no notion of the number of dimensions and their units of measurement. Resnik (Resnik and Diab, 2000) refers to human participants finding similarities of meaning according to dimensions that we have not yet formalised. Vigliocco (Vigliocco et al., 2002a) however, reports operationalising semantic distance as the Euclidian distance between the two best responding input units in a feature space encoded in a Kohonen self-organising map.

1.4.3.2 The Vector Space Model

Salton's Vector Space Model (VSM) described a method for clustering documents in a semantic space based on vectors of extracted terms (Salton et al., 1975). "Term" can mean a single word or a contiguous phrase – for example Random Access Memory. It also introduced a systematic method of identifying good terms, weighting terms based on their

importance in a particular ontology and combining poor terms to synthesise better terms. The model incorporates, Spärck-Jones' (Spärck-Jones, 1972) TD/IDF approach.

The VSM continues to influence many developments in IR where textual items are presented as unit vectors and similarity is calculated as the cosine of the angle between them (Cardie et al., 2000), (Erkan and Radev, 2004), (Kimura et al., 2007), (Yeh et al., 2008), (Saggion et al., 2003), (Hatzivassiloglou et al., 2001, Park et al., 2000).

Under the Vector Space model, each text (document) D_i, is represented by indexing terms T*j* which are weighted according to their importance in indexing the document, resulting in a vector which represents the document (equation 1.4):

$$D_i = [d_{i1}, d_{i2}, d_{i3} \ldots d_{it}] \tag{1.4}$$

where the document has *t* indexing terms and d_{ij} represents the weight of the *j*th term.

In the simplest case the weights can be restricted to 0 or 1 representing the absence of presence of the term in the document.

The Vector Space model uses the TF*IDF term weighting method (Spärck-Jones, 1972) for calculating the weights used in the vectors (equation 1.5):

$$TF * IDF = \mathrm{TF}[word_i] \cdot \log \frac{TNU}{DF[word_i]} \tag{1.5}$$

Where TF = total frequency of words in each text unit, TNU = total number of text units, DF = document frequency (the total number of text units each word appears in) (Hatzivassiloglou et al., 2001)

Salton specified the inner product or cosine measure as the way of computing the similarity between the two vectors. This may be computed by equation 1.6:

$$Sim[D_i, D_j] = \frac{|D_i \bullet D_j|}{|D_i||D_j|} \tag{1.6}$$

Although the VSM has been used as the basis of measuring ST similarity, it suffers from a serious weakness – it is a lexical measure. The elements of the vector concern words as symbols and do not have any knowledge of their meanings. At best, operations such as stemming or lemmatisation (Manning et al., 2008) are used to reduce inflections of a particular word to a single form. Such measures have been criticised as working poorly when there is little word overlap (Jeon et al., 2005).

Other disadvantages of this technique for sentence measures are inefficiency (many null entries in the vectors) and exclusion of useful function words (e.g. *this, not* etc.)

1.4.4 Fundamental Models of Short Text Semantic Similarity, STASIS and LSA

The bulk of the work described in this thesis concerns the creation of datasets for the evaluation of STSS measures. At the inception of the work there were only two measures which were clearly and intentionally useful for STSS, STASIS and LSA.

Bearing in mind the need to evaluate the datasets (or at least illustrate their use), STASIS and LSA were chosen for use throughout this thesis.

Following the larger-scale publication of STASIS (Li et al., 2006), there has been a flurry of work in the STSS field. Virtually all of this work has been inspired in some way by either STASIS (combining word similarities derived from WordNet path lengths, using short word vectors and joint word sets, or using word order information) or LSA (by actually incorporating LSA as a component).

In fact, by far the majority of subsequent work in the field is either derivative from or influenced by STASIS (Ferri et al., 2007, Tsatsaronis et al., 2010, Min et al., 2006, Gacitua-Decar and Pahl, Huang et al., 2008, Capuano et al., 2009 , Inkpen, 2007, Achananuparp et al., 2008, O'Shea et al., 2009, O'Shea et al., 2008, Liu et al., 2007a) or LSA (Tsatsaronis et al., 2010, Inkpen, 2007, Feng et al., 2008, Pilato et al., 2009, Augello et al., 2008) in some way. This endorses the decision to use STASIS (the most seminal) and LSA (the most different from STASIS) as vehicles for demonstrating the evaluation and comparison processes. Other processes used in STSS measures such as string matching stray away from true semantic similarity into surface lexical similarity (Stoilos et al., 2005).

Due to the importance of these two models to the work in this thesis, their algorithms are discussed in detail in the following two sections.

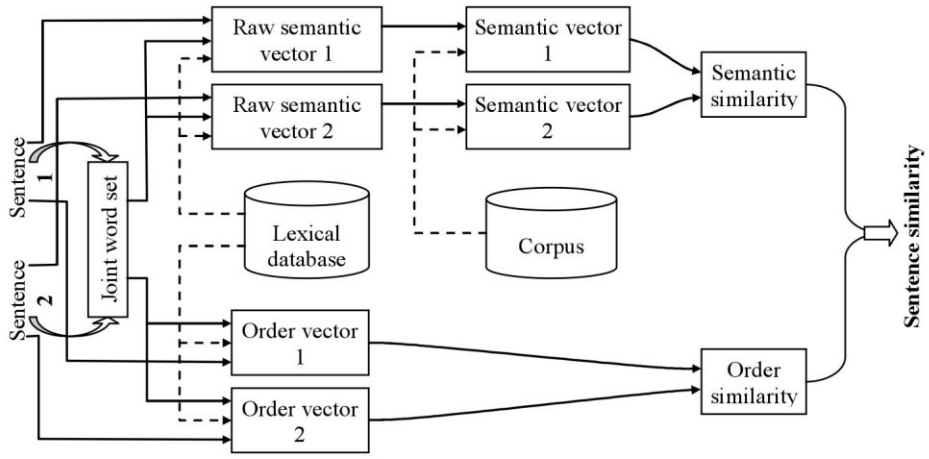
1.4.4.1 STASIS

STASIS (Li et al., 2006) was specifically designed to overcome the problems of high dimensional vector-space models. Its key features are: short vectors derived only from the words in the STs, use of function words, specific word forms (no stemming/lemmatisation) and exploitation of word order information. Function words are high-frequency closed-class words e.g. articles and auxiliary verbs. In the two sentences "Could you pass the salt?" and "Did you pass the salt?" a single word changes the speech act (Austin, 1975), the overt meaning and the subtle implications of the basic propositional content.

Figure 1.2 shows the two stages used to calculate the overall semantic similarity: construction of two vectors (semantic and word-order), followed by combination of the similarity information obtained by the vectors. In the following examples, taken from (Li

et al., 2006), the lexical database, corpus and word similarity measure components can be replaced by alternatives, although the authors recommend using their own word similarity measure (Li et al., 2003).

Figure 1.2 STASIS sentence similarity computation diagram



Construction of the Joint Word Set

Equation 1.7 describes a joint word set T derived from all of the distinct words in two short texts, T1 and T2.

$$T = T_1 \cup T_2 = \{w_1 \quad w_2 \quad \cdots \quad w_m\} \tag{1.7}$$

For example, given T1 and T2:

T1: RAM keeps things being worked with.
T2: The CPU uses RAM as a short-term memory store.

The joint word set, T is:

{RAM keeps things being worked with The CPU uses as a short-term memory store}

Formation of the Lexical Semantic Vectors

A lexical semantic vector $\check{s}$ is derived from the joint word set for each short text. The dimensionality m equals the number of words in the joint word set. Each entry, $\check{s}_i(i=1,2,...,m)$, is determined by the semantic similarity of the corresponding word in the joint word set to a word in the short text.

The joint word set is scanned from beginning to end and one of two possible actions is taken for each word

Case 1:If w_i appears in the short text, $\check{s}_i$ is set to 1.

Case 2:If w_i is not contained in T_1, a semantic similarity score is computed between w_i and each word in the short text T_1, using the method described in (Li et al., 2003).

The most similar word in T_1 to w_i is that with the highest similarity score ς. If ς exceeds a preset threshold, then $\check{s}_i = ς$, otherwise $\check{s}_i = 0$.

The threshold is used because it is assumed that below this level the values are merely contributing noise.

Equation 1.8 shows how the words are then weighted according to their information content (Resnik, 1999), on the assumption that word frequency influences the contribution of the individual words to the overall similarity. Entropy measures are calculated using the Brown corpus, although other corpora such as BNC could be used if desired.

$$s_i = \check{s} \cdot I(w_i) \cdot I(\tilde{w}_i) \tag{1.8}$$

Where $I(w_i)$ is the information content of a word in the joint word set and $I(\tilde{w}_i)$ is the information content of its associated word in the short text. Therefore both words make distinct contributions to the short text similarity based on their individual information contents.

Calculation of the Semantic Similarity component

Having generated the semantic vectors, s1 and s2, the semantic similarity (S_s) between the two short texts can be calculated using a cosine-like measure between the two vectors, using equation 1.9:

$$S_s = \frac{\mathbf{s}_1 \cdot \mathbf{s}_2}{\|\mathbf{s}_1\| \times \|\mathbf{s}_2\|} \tag{1.9}$$

Formation of the Word Order Vectors

The following sentence pair is an extreme case illustrating the importance of word order:

T_1: A quick brown dog jumps over the lazy fox.

T_2: A quick brown fox jumps over the lazy dog.

Any measure which uses words alone without taking account of position will classify them as identical in meaning, yet this is clearly not the case. The joint word set for the two sentences is:

T = {A quick brown dog jumps over the lazy fox}

Each word in T_1 and T_2 is assigned the order number that the word appears in the sentence. For example, in T_1 the index number is 4 for *dog* and 6 for *over*. A word order vector is generated from the joint word set for each short text T_1, T_2 ($\mathbf{r_1}$, $\mathbf{r_2}$ respectively).
For example, $\mathbf{r_1}$ is constructed by scanning the joint word set from beginning to end and taking one of two possible actions for each word.

Case 1: If w_i appears in T_1, the corresponding word $\mathbf{r_{1j}}$ is set to the index of w_i in T_1

Case 2: If w_i is not contained in T_1, a semantic similarity score is computed

between w_i and each word in the short text T_1 to find the most similar word $\tilde{w}_i$ with a similarity score ς. If ς exceeds a preset threshold, then the entry of w_i in $\mathbf{r}_1$ is filled with the index number of $\tilde{w}_i$ in T_1, otherwise it is set to 0. Applying this in the example produces the vectors:

$\mathbf{r}_1 = \{1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\}$
$\mathbf{r}_2 = \{1\ 2\ 3\ 9\ 5\ 6\ 7\ 8\ 4\}$

Calculation of the Word Order Similarity component

Equation 1.10 shows how word order similarity is (S_r) calculated as the normalised difference in word order:

$$S_r = 1 - \frac{\|\mathbf{r}_1 - \mathbf{r}_2\|}{\|\mathbf{r}_1 + \mathbf{r}_2\|} \qquad (1.10)$$

Thus it takes account of the number of shared words, their order, the distance between them and the overall length of the texts. A rationale for this metric with a worked example is provided in (Li et al., 2006).

Calculation of Overall Sentence Similarity

The overall similarity of the two sentences is determined by adding the semantic similarity component and the word order component with a desired range of 0 to +1. Both components produce values in this range and the word order component is considered to contribute less than the semantic similarity component and the desired range. Therefore the components are combined using equation 1.11:

$$S(T_1, T_2) = \delta S_s + (1 - \delta) S_r \tag{1.11}$$

Substituting for S_s and S_r gives equation 1.12:

$$S(T_1, T_2) = \delta \frac{\mathbf{s}_1 \cdot \mathbf{s}_2}{\|\mathbf{s}_1\| \times \|\mathbf{s}_2\|} + (1 - \delta) \frac{\|\mathbf{r}_1 - \mathbf{r}_2\|}{\|\mathbf{r}_1 + \mathbf{r}_2\|} \tag{1.12}$$

The parameter δ (which adjusts the relative contributions of semantic and word order) is in the range $0.5 < \delta < 1$; α, β and δ are all chosen empirically.

1.4.4.2 LSA

Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text (Landauer et al., 1998). Although it is designed for document retrieval, one by-product is a facility for submitting sentences to the LSA website for similarity measurement (Laham, 1998). LSA requires two stages, construction of semantic space, then calculation of the cosine of the vectors representing the pair of STs under consideration in the space.

Construction of the high dimensionality Semantic Space

Once the space has been created it can be re-used repeatedly to process queries. There are a number of stages involved in construction. The following example is taken from (Landauer et al., 1998) and (Deerwester et al., 1989).

The text is pre-processed to identify potentially useful words and phrases using cues such as punctuation marks, function words and parenthetical phrases. This is followed by stemming to reduce word variants into a single semantic form, using heuristics with exception lists.

The words are used to compile a lexicon and parsing is applied to discover potentially useful phrases. Function words are stripped out and a term-by document matrix is constructed. Terms that occur in more than one document appear as rows in the matrix and the documents themselves appear as columns. Each cell in the matrix is a count of the number of times the term occurs in the particular document. Table 1.1 shows a small matrix built using 12 terms from 9 documents.

In this case the "documents" are titles, (short texts). For example C1: "Human machine interface for Lab ABC computer applications" is responsible for the value of 1 entered in the cells for Human, Interface and Computer in column C1. Shannon's entropy measure (equation 1.13) is used to weight the counts in the cells, followed by normalisation.

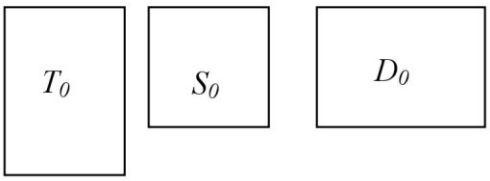
$$H[p] = -p\log[p] \tag{1.13}$$

	Documents								
Terms	C1	C2	C3	C4	C5	M1	M2	M3	M4
Human	1	0	0	1	0	0	0	0	0
Interface	1	0	1	0	0	0	0	0	0
Computer	1	1	0	0	0	0	0	0	0
User	0	1	1	0	1	0	0	0	0
System	0	1	1	2	0	0	0	0	0
Response	0	1	0	0	1	0	0	0	0
Time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
Survey	0	1	0	0	0	0	0	0	1
Trees	0	0	0	0	0	1	1	1	0
Graph	0	0	0	0	0	0	1	1	1
Minors	0	0	0	0	0	0	0	1	1

Table 1.1 Term by document matrix

This large matrix is decomposed by Singular Value Decomposition (equation 1.14) into 3 other matrices:

$$X = T_0 S_0 D_0' \tag{1.14}$$



S_0 is a diagonal matrix of dimensionality m. At this stage the original term-document matrix which has a dimensionality (t,d) can be reconstructed perfectly from the three new matrices. Matrix T_0 has dimensionality (t,m) and matrix D_0 has dimensionality (m,d) and m = min(t,d). The variable m represents the dimensions in the constructed semantic space. T_0 contains the terms as rows and the semantic space dimensions as columns. D_0 contains the documents as rows and the semantic space dimensions as columns. Therefore both the terms and documents can be represented as vectors in the same m-dimensional semantic space. The next step is reduction of the dimensionality. The diagonal matrix S_0 has dimensionality (m,m) (all non-diagonal elements are 0). As a result of SVD the members of S_0 are ordered by size (table 1.2).

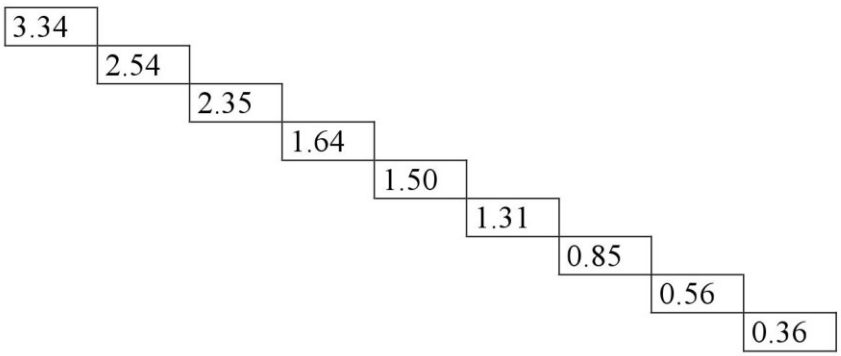


Table 1.2 Diagonal matrix

Small elements of the diagonal matrix can be set to zero, then deleted with corresponding reductions in the size of T_0 and D_0. The *k* largest elements remain unchanged allowing an approximation of the original matrix to be reconstructed (equation 1.15).

$$X \approx \hat{X} = TSD' \qquad (1.15)$$

Setting a threshold of 2.5 retains the first 2 elements of the diagonal matrix (k=*2*) resulting in a 2-dimensional semantic space shown in table 1.3. The generalising power of LSA can be understood by examining the differences between X and $\hat{X}$. Some cells increase in value and some decrease. For example a cell with a value of zero (no occurrence of the term in the document) may now contain a positive number (indicating that in the full population of documents it should occur to some extent in this type of document).

	Dimensions	
Terms		
Human	0.22	-0.11
Interface	0.20	-0.07
Computer	0.24	0.04
User	0.40	0.06
System	0.64	-0.17
Response	0.27	0.11
Time	0.27	0.11
EPS	0.3	-0.14
Survey	0.21	-0.27
Trees	0.01	0.49
Graph	0.04	0.62
Minors	0.03	0.45

Table 1.3 Diagonal matrix reduced to 2 dimensions

So information about the terms is generalised and spread from one cell into a number of related cells, which allows LSA to calculate a similarity even when no co-occurrence exists between particular terms.

Similarity measurement

The term-with-term feature can be used to compare two short texts. A row in the semantic space is formed for each short text. It is important to be able to measure similarity of texts which did not appear in the original. A method for constructing a Pseudo-Object to do this is disclosed in an LSA patent (4,839,853). The method involves identifying the terms from the text which occurred in the original matrix X, multiplying by the term vectors and then by the inverse of the singular values as in equation 1.16:

$$D_q = M_q^T TS^{-1} \tag{1.16}$$

Once each text has a row vector in $\hat{X}$ the cosine of the angle between these vectors (dot product) is a measure of their similarity (equation 1.17):

$$\hat{X}\hat{X}' = TS^2T' \tag{1.17}$$

LSA's performance is highly dependent on the choice of the parameter *k* and the corpus used to generate the semantic space. *k* may be chosen empirically (Deerwester et al., 1990) or through some external criterion of validity (Landauer et al., 1998) and is typically in the range 50 – 1500. All of the examples on the LSA website use the corpus General Reading Up to 1st Year College (with *k*=300 as their default setting).

This corpus consists of 37,651 different documents extracted from various textbooks, novels, newspaper articles, and other information. After statistical analysis 92,409 terms were identified and the semantic space contains 419 dimensions.

1.4.5 Evaluation: Agents vs. Algorithms

There has been an interest in evaluating dialogue-based systems (in some form or other) since the late 1990s. Although these early attempts provide an interesting background, this section establishes that the applications themselves were so simple that there was little to actually evaluate.

Early developments were largely driven by the need to evaluate Spoken Dialogue Systems (SDSs) under development at the time (Lamel et al., 1997, Gorin et al., 1997, Bouwman and Hulstijn, 1998, Bouwman et al., 1999).

The applications at the time were quite trivial - for example, PADIS (Bouwman and Hulstijn, 1998) routed a telephone call arriving at a switchboard to the correct extension or retrieved other information such as an employee's e-mail address. Other systems were concerned with simple booking systems, for example for train tickets (Lamel et al., 1997, Lamel et al., 2002). The main challenge in SDS systems was the accuracy of speech recognition so it is unsurprising that the tasks dependent on this challenging component were relatively simple. The dialogue from PADIS (Philips Automatic Directory Information System) is a good example:

S1: PADIS. What can I do for you?
C1: I'd like to talk to Mike.
S2: So, would you like to talk to Mike Anderson?
C2: Yes, please.
S3: The extension 506 is busy right now. Can I do anything else for you?
C3: Oh yes. Can you give me the e-mail address of Mary, please?
S4: Which Mary do you want the e-mail address of?
C4: Uh, she is in the speech recognition group.
S5: So you want the e-mail address of Mary from the speech recognition group?
C5: Exactly.
S6: The e-mail address is . . .

In this system the employees were identified by name, title, gender or working group and the dialogue shows the continuous checking of the information extracted from speech recognition, because of the likelihood of errors. The characteristic features of such systems are the simplicity of the task, the shortness of the transactions and the extreme cooperativeness of the human participants.

Since the year 2000, there has been a growing interest in developing (and therefore evaluating) embodied conversational agents (Cassell et al., 2000). The MIT media lab is a centre of work on embodied CAs (Bickmore and Cassell, 2000, Cassell et al., 2001, Cassell and Bickmore, 2003). ECAs in fact, do little to advance the field of dialogue. For example this dialogue:

1. How about this weather?
2. I think winters in Boston are awful.
3. How do you like Boston?
4. Boston is certainly more expensive than it used to be.
5. So where would you like to live?
6. How many bedrooms do you need?

Etc.

illustrates the use of chatterbot-style social engagement combined with a linear attempt to collect a very small number of attributes in order to make a property recommendation. The

main focus of the paper it came from was an investigation of whether or not incorporating social chat could foster trust in the agent by the user (Bickmore and Cassell, 2000).
Taking embodiment a stage further, gestures on the part of the agent, and recognition of gestures by the user tend to make up for shortfalls in the actual process of dialogue management, for example (user Tim, Agent Rea):

> Tim says: "Show me the kitchen."
> Rea shifts the viewpoint to show the interior of the kitchen, and says "It is a modern kitchen."
> Tim says "I like blue tiles." With a beat gesture on the word 'blue'.
> Rea responds by saying "Blue is my favourite color."
> Tim says "I like blue tiles." With a beat gesture on the word 'tiles'.
> Rea responds by saying "I love tiles."

In this passage, a beat gesture is "a small, rhythmic baton-like movement of the hands" indicating what is new in the speaker's utterance (Cassell et al., 2001). Despite the improvements in understanding of utterances that may come from the multi-modal operation, the challenge of the tasks remained straightforward and basically linear in nature.
By the middle of the last decade some more challenging application areas were being attempted, for example advising users of digital document collections (Andersen et al., 2003), Intelligent Tutoring (Graesser et al., 2005), training children with hearing loss (Massaro et al., 2000) and museum and gallery and other guides (Michie, 2001, Kopp et al., 2005, Yuan and Chee, 2005, Le Bigot et al., 2004).
More recent developments have, by and large, failed to move on to realistically challenging applications. There has been further work on museum and tour guides (Robinson et al., 2006, Damiano et al., 2008, Misu and Kawahara, 2010, Thomson et al., 2008) and virtual receptionists (Babu et al., 2006, Robinson et al., 2006).
There has also been an interest in the "softer" aspects of user-agent interaction. Interest in affective behaviour dates back to 2003 (Mori et al., 2003), where galvanic skin resistance and blood pressure were monitored by users playing a mathematical game with an agent and different verbal and nonverbal behaviours on the part of the agent were investigated to try and reduce users' negative emotions.
Related work includes evaluation of the effect of two different non-verbal behaviours on users' emotional reactions (Krämer et al., 2007) and investigation of the agent as a social entity (Babu et al., 2006). There was also research interest in applying conversational agents to e-commerce in the early half of the last decade (Semeraro et al., 2003), which has persisted (Semeraro et al., 2008, Matthews et al., 2008, Pu and Chen, 2006). However,

reviews of recent work suggest that the scope of avatar-based selling interfaces remains limited (McGoldrick et al., 2008).

A few exceptional systems address more serious topics. Goh et al. (Goh et al., 2007) described a 5-layer architecture embodied CA to provide information about bird flu. Despite its complexity (and heavy use of NLP processes) this system is basically a question answering system, whose contribution is the searching and understanding of web sources to find answers to user questions. This is illustrated in the following extract:

> Isabel_D1 : Have you heard anything about bird flu?
> AINI _D1 : Did you mean H5N1 avian flu virus, a global flu pandemic?
> Isabel_D2 : Yes.
> AINI _D2 : Why do you ask?
> Isabel_E1 : Just want to know what is bird flu pandemic?
> AINI _E1 : Bird flu or avian influenza is an infectious disease of birds ranging from mild to severe form of illness. Some forms of bird flu can cause illness to humans.
> Isabel_F1 : Can a pandemic be averted and how many has been confirmed?
> AINI _F1 : Cumulative number of confirmed human cases of avian influenza A/(H5N1) reported by WHO. More information at http://www.who.int/csr/disease/avian_influenza/country/cases_table_2006_03_21/en/index.html
> Isabel_F2 : Thanks for your information, can we talk about our future.
> AINI _F2 : My pleasure. I'm not sure how this related to Bird Flu? Why don't you ask me about How is bird flu transmitted to humans?

However, the system does not appear to have any diagnostic capability, even that required by the simple checklist used in the UK by NHS direct for authorising the issue of tamiflu. Given the complexity of this system it would be totally incapable of scaling up to the real-world pressures experienced by the NHS direct website (BBC, 2009).

A review by Bickmore (Bickmore and Giorgino, 2006), which reports positively on the use of the pattern matching technique, concludes that "health information and counselling using natural language dialog . . . in some cases, are as effective as human health providers" (although providing information and counselling largely boils down to relatively straightforward question answering and Eliza-like techniques).

Some challenging applications have been tackled using the CA architecture shown in figure 1.1. These include advising employees on bullying and harassment procedures and leading them through the various stages of complaint required to end the behaviour (Crockett et al., 2009). A second example reported in the same paper is assistance and advice on student debt, which crucially leads the student through various alternative processes required to pay off the debt. A third example, so far unreported (from the same research group) leads people who wish to bid for a social housing property through the necessary procedures. The common feature across these applications is that they are high

stakes and may involve the users engaging in deceptive, manipulative, adversarial or even confrontational behaviour with the agent to achieve an end.

1.4.5.1 Holistic Evaluation of Agents

A substantial amount of work has been done on evaluating agents as a whole. These break down into 3 categories:

- Aspirational Subjective Measures
- Attempted Subjective Measures
- Objective Measures

1.4.5.2 Aspirational Subjective Measures

A number of publications discuss very high level, abstract and subjective concepts which would be very difficult to measure as a single attribute. The most common attributes are:

- Usability (Bouwman et al., 1999, Semeraro et al., 2003, Andersen et al., 2003, Lamel et al., 2002, Walker et al.)
- User satisfaction (Bouwman et al., 1999, Yuan and Chee, 2005, Lamel et al., 2002, McKevitt et al., 1999, Le Bigot et al., 2004)
- Agent credibility (Yuan and Chee, 2005, Cassell and Vilhjálmsson, 1999, Massaro et al., 2000)

The first two are common (if difficult) attributes measured across the field of software engineering (Fenton and Pfleeger, 1998). There are many more intangible and vague attributes mentioned in studies, including: "Fun to talk with" (Kopp et al., 2005), "lovely, pleasant, black humorous" (Yuan and Chee, 2005), "Intimacy, Benevolence" (Bickmore and Cassell, 2000), "Comfort, Solidarity, Familiarity" (Cassell and Bickmore, 2003) and "Trust, Uncertainty, Attractive" (Semeraro et al., 2003).

1.4.5.3 Attempted Subjective Measures

Some studies then go on to attempt to measure a subset of subjective attributes. These are largely measured using Likert or Likert-like attitude rating scales. Attributes measured in this way include:

- Ease of use / Task ease (Lamel et al., 1997, Litman and Pan, 2002, Sanders and Scholtz, 2000, Walker et al., 2001)

- Ease of the user understanding the agent (Litman and Pan, 2002, Semeraro et al., 2003, Lamel et al., 1997, Walker et al., 2001)
- The agent's understanding of the user comprehension (Litman and Pan, 2002, Semeraro et al., 2003, Bouwman et al., 1999).
- Various cognitive attributes related to comprehension and complexity (Cassell and Vilhjálmsson, 1999, Andersen et al., 2003, Lamel et al., 1997, Bouwman et al., 1999)
- Various attributes related to the reliability of the agent and the ease of correcting misunderstandings (Bouwman et al., 1999, Lamel et al., 1997)
- Various attributes concerning the user's expertise (of the domain or using the agent) (Litman and Pan, 2002, Semeraro et al., 2003, Andersen et al., 2003, Walker et al., 2001)
- The efficiency or effectiveness of the agent (Sanders and Scholtz, 2000, Lamel et al., 1997, Semeraro et al., 2003, Andersen et al., 2003, Bouwman et al., 1999)
- Various attributes about command and control of the conversation (Cassell and Vilhjálmsson, 1999, Lamel et al., 1997, Bouwman et al., 1999) (Semeraro et al., 2003, Andersen et al., 2003)
- The pace of the interaction (Litman and Pan, 2002, Lamel et al., 1997)
- Whether the agent behaved as expected (Litman and Pan, 2002, Bouwman et al., 1999, Walker et al., 2001)
- How natural the agent's behaviour seemed (Cassell and Vilhjálmsson, 1999, Cassell and Bickmore, 2003, Sanders and Scholtz, 2000)
- Various positive emotional attributes (e.g. friendliness, enjoyment) (Semeraro et al., 2003, Bouwman et al., 1999) (Cassell and Bickmore, 2003, Lamel et al., 1997)
- Various negative emotional attributes (e.g. boredom, fluster) (Bouwman et al., 1999, Lamel et al., 1997)
- Whether the user would use again (Litman and Pan, 2002, Bouwman et al., 1999, Lamel et al., 1997, Walker et al., 2001) or prefer human service (Lamel et al., 1997)

There were also a substantial number of attributes which occurred once or twice including "like further help" (Bouwman et al., 1999), "narrative skills" (Yuan and Chee, 2005), "needs improvement" (Lamel et al., 1997), "question answering capability" (Yuan and Chee, 2005) and "how much willing to pay" (Bickmore and Cassell, 2000).

1.4.5.4 Objective measures

Most studies include a set of objective measures. Generally speaking, there is a leap of faith that these in some way reflect the aspirational subjective measures that appear at the beginning of published studies. The only systematic and scientific approach was that taken by the Paradise framework (Walker et al., 2001). Attributes measured in this way include:

- Dialogue / Conversation length (Bouwman et al., 1999, Cassell and Vilhjálmsson, 1999, Kopp et al., 2005, Walker et al., 2000, Walker et al., 2001, Le Bigot et al., 2004)
- Count of dialogue turns (Litman and Pan, 2002, Bouwman and Hulstijn, 1998, Graesser et al., 2005, Le Bigot et al., 2004, Walker et al., 2001, Walker et al., 2000, Sanders and Scholtz, 2000)
- Various measures of success at utterance or task completion level (Bouwman et al., 1999, Bouwman and Hulstijn, 1998, Lamel et al., 2002, Le Bigot et al., 2004, Walker et al., 2001)
- Various counts of errors, corrections or percentage error rates (Litman and Pan, 2002, Kopp et al., 2005, Bouwman and Hulstijn, 1998, McKevitt et al., 1999, Bickmore and Giorgino, 2006)
- Various counts of correct actions by the agent (e.g. answering questions) (Yuan and Chee, 2005, Semeraro et al., 2003, Andersen et al., 2003, Walker et al., 2001)
- Various speech recognition accuracy measures (Massaro et al., 2000, Graesser et al., 2005, Lamel et al., 1997)

There were also a substantial number of attributes which occurred once or twice including "mental workload" (Le Bigot et al., 2004), "learning gains" (in a tutoring system) (Graesser et al., 2005), count of help messages (Lamel et al., 1997), percentage of time user spent looking at (embodied) agent (Andersen et al., 2003) and user trust of agent (using a standardised measure from psychology) (Bickmore and Cassell, 2000).

The most comprehensive set of metrics devised for CA evaluation is the Paradise framework (Walker et al., 2001). This framework applied the process generally used in Software Engineering (Fenton and Pfleeger, 1998) for deriving abstract, indirect attributes such as User Satisfaction in terms of directly measurable attributes, using linear regression.

1.4.5.5. Evaluation of algorithms

Despite all of the work described in section 1.4.5.4, black box measures of integrated conversational agent systems do not provide much insight into the performance of a new algorithm forming a component of a CA.

Suppose we wished to assess the performance of STSS vs. Pattern Matching in a CA. We might build two systems using the two approaches and compare them on subjective measures such as *The agent's understanding of the user* or objective measures such as *Count of dialogue turns*. However, using the two different approaches may lead to other differences between the two agents during the build process. This is even more pertinent in cases where agents developed by different research groups, using different algorithms, are compared.

The *count of dialogue turns* may in itself be a component contributing to the indirect measure *user understanding*, because poor understanding will lead to the user going down blind allies, requiring corrective turns and also requests for clarification which will also add turns to the conversation. Therefore any additional differences that lead to lack of understanding will confound the comparison of the two algorithms embedded in the agents (STSS vs. Pattern Matching). Specifically, differences in the verbal content of the responses by the agent to the user, differences in the underlying model of the problem domain and differences in the interface design may contribute confounding factors.

The consequence is that when a new algorithm such as STSS is proposed for incorporation into CAs it must be thoroughly evaluated in isolation first (unit testing), to do otherwise would be building on sand. Proper evaluation requires the use of appropriate statistical methods, the creation of standard benchmark datasets and a sound understanding of the properties of such datasets. This position is supported by work by (Mihalcea et al., 2006) who evaluated 2 corpus-based and 6 knowledge-based word similarity measures before incorporating one in a text similarity measure.

Because STSS is a novel approach, there are no established methods for evaluating such measures. Also, because semantic similarity is characterised by human perception there is no "ground truth" similarity rating that can be assigned to pairs of STs, the only way to obtain ratings is through carefully constructed experiments with human participants.

1.4.5.6 Evaluating STSS measures

The evaluation of STSS measures is covered in depth in chapters 2 and 3. The two key questions, which must be addressed to evaluate STSS, are:

(i) "What constitutes a correct semantic similarity assessment?"

(ii) "What is a suitably representative data set?"

Experiments which require human ratings of psychological attributes such as similarity largely depend upon techniques devised by psychologists for measuring human attitudes in social sciences studies.

A crucial element in developing the measurement instruments (questionnaires etc.) and procedures used in these studies is piloting (Oppenheim, 1992).

Given the difficulties involved in obtaining the ground truth in experiments capturing subjective ratings, the pilot study (chapter 2) focused on devising the best experimental techniques for obtaining the human ratings. Although the data set that was generated in chapter 2 was restricted to a single type of dialogue act, it has been adopted as a Gold Standard by researchers in the field. Chapter 3 set out to produce a more demanding data set, which will have long-term value in demonstrating improvements in CAs.

1.5 Conclusions

An important part of reaching the conclusions of this chapter is to acknowledge a debt to history. Since the publication of the original STASIS measure of STSS (Li et al., 2004, Li et al., 2006), a number of new STSS measures have emerged (work reviewed in later chapters of this thesis). Consequently, during this work there has been a constant awareness of the need to find new, computationally efficient methods of improving STSS measurement.

Much additional reading of the literature took place during the early phases of this work which are not appropriate to report in depth. However, one chain of research must be acknowledged. Beginning with an interest in the seminal work of Alan Turing (Turing, 1950), this led to later work by philosophers intent on debunking the Turing Test (Block, 1981), (Searle, 1980). Extended reading of the work of Searle led to the concept of the speech act (Searle, 1969), and the equivalent Dialogue Act in the field of dialogue management. This provided the inspiration for the work in chapters 4 and 5 on a factor–base model in which the dialogue act is separated out from propositional content in a user utterance.

The other chain of reading triggered by Turing's work led backwards to the Cartesian test for automatism (Erion, 2001) based on the *Traite de l'homme* of René Descartes (1596-1650), modern translation published in (Descartes, 1988). The philosophical chain also led to the work of Gottfried Wilhelm von Leibniz (1646 – 1716), whose work led to the

formulation of Leibniz' Law, the Identity of Indiscernibles. This law states that two things are identical if and only if they share the same and only the same properties. This may be the first feature-based approach to similarity, leading to the work of Amos Tversky which is a major influence on current views of similarity. This feature-based strand of similarity work is practically embodied in Signal Detection Theory (Egan, 1975), used in chapter 3 of this thesis as an alternative measure to validate the ratings obtained for the new gold-standard STSS dataset. So although the work of Alan Turing makes no explicit contribution to the work described in this thesis, it has been profoundly influential in fostering a deeper understanding of the field and the interconnected strands of philosophy, psychology and neuroscience that permeate the work.

What is the outcome of the diverse reading of the literature reported in this chapter? The evidence from the review of existing CAs supports the conclusion that no method is going to emerge from their supporting technologies that will allow rapid, economical and maintainable CAs to be produced on a commercial basis in the forseeable future. An important distinction was made between the general field of CAs and Goal-Oriented CAs. The general field of CAs is broad, encompassing diverse elements such as Embodied CAs and telephone Automatic Speech Recognition systems. This general approach lacks the focus, required to solve difficult real-world problems, that is provided in Goal-Oriented CAs. However, despite their improved strategy, Goal-Oriented CAs inherit some of the problems faced by the general class of CAs, therefore a paradigm shift is required to make progress.

STSS is a promising technique but much work will be needed to develop suitable algorithms, embed them in CAs and perform the necessary stages of evaluation.

The broadest possible coverage was given to the fields of similarity (in general) and semantic similarity (in particular).

This was for two reasons. The first was to establish the credibility of similarity as a fundamental attribute emerging from human experience of the world, which can also be treated as a credible scientific variable, used in properly designed experiments and amenable (to a reasonable degree) to numerical processing, statistical analysis etc. The second was a reason ingrained in any scientific researcher – to determine whether the topic being investigated has also been investigated by another branch of science and discover whether experimental questions have been answered, problems solved or alternative techniques developed that could be built on by the proposed work. Above all, to ensure that the proposed contribution to human knowledge is indeed original in the wider context. The review of various aspects of cognitive science proved particularly fruitful in solving a

problem discussed in chapter 3, which had appeared intractable at the commencement of the work in this thesis.

The evidence from the literature search also supported the position that similarity is a human attribute that exists across a wide range of human experience and is relied on for high-stakes real-world applications such as DNA matching. Furthermore, that although more difficult to deal with than physical attributes such as mass, voltage and temperature, it is capable of being put on a quantitative basis to a sufficient degree to support the use of numerical measurement and processing by computer algorithms.

The literature also highlighted the importance of developing sound benchmark datasets for the evaluation of similarity, and a good basis for developing STSS benchmark datasets was found in prior work on word semantic similarity. It also revealed the importance of the Gold Standard dataset, one which is produced using the best possible available techniques and which becomes adopted as the standard by which future improvements are tested.

Finally, as stated in the opening remarks of these conclusions, the themes followed in the literature led to the recognition of the fundamental importance of the Dialogue Act in measuring similarity between short texts.

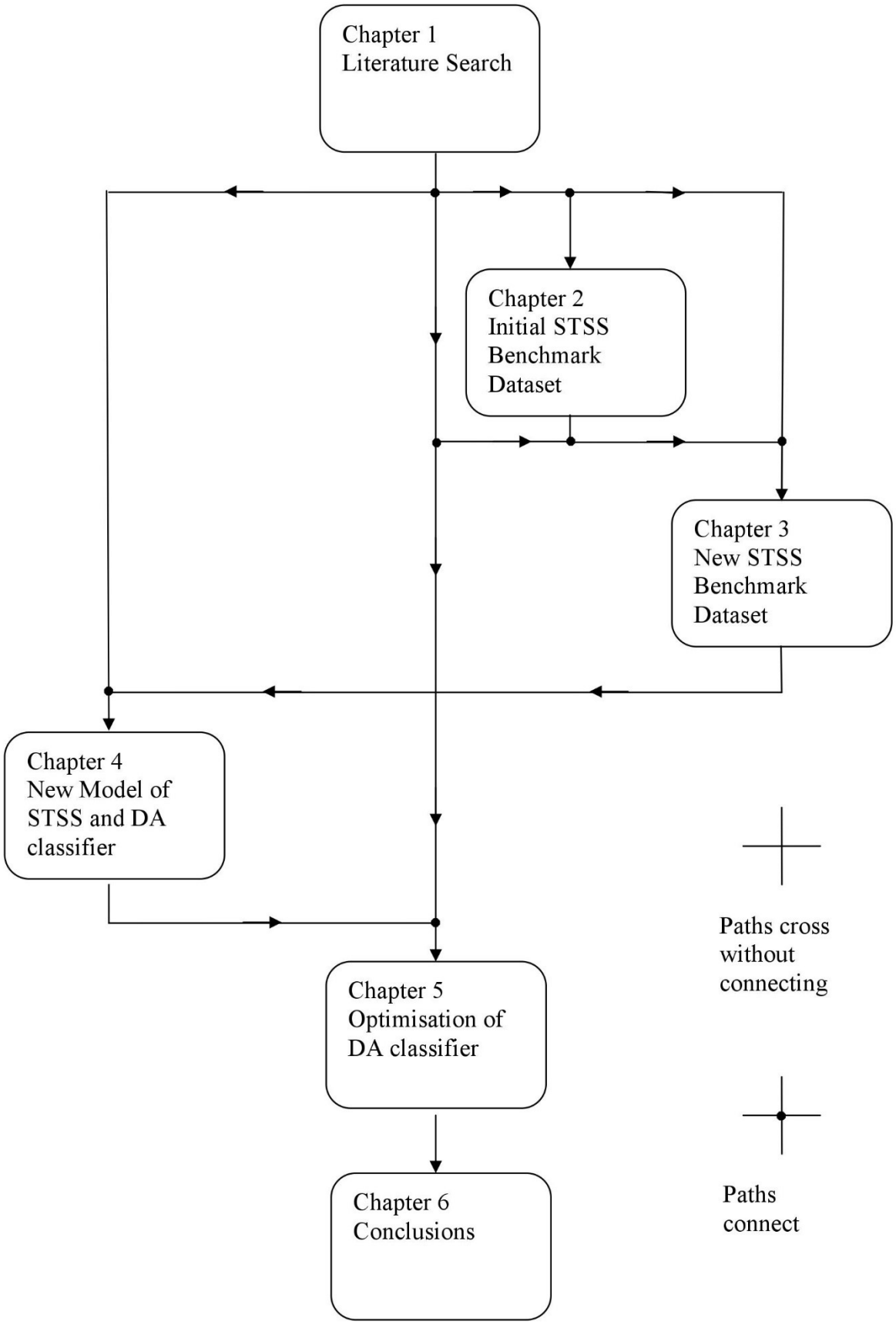
1.5.1 Relationships between chapters

Figure 1.3 outlines the relationships between the chapters comprising the thesis. The literature search in chapter 1 investigates the role and properties of the attribute of similarity across a wide range of disciplines. This establishes that the problems addressed by this thesis have not been solved in other disciplines. It also provides the research background to support Chapter 2 (material on word similarity, general similarity theory), Chapter 3 (social sciences research methodology, Category Specific Deficits) and chapter 4 (Dialogue Acts)

Chapter 2 provides the first data set for benchmarking short text semantic similarity and also supports the work of Chapter 3 by investigating the best experimental methods for obtaining human similarity ratings. This is achieved by using a limited, but well studied (in terms of semantic similarity), set of words as the source of sentences for the experiments.

Chapter 3 provides a full benchmark data set which aspires to be representative of the English language as a whole. The codification of procedures in chapter 2 allows Chapter 3 to focus on representation, for example by promoting the use of diverse dialogue acts. The experience gained with Dialogue acts in this chapter again supports the future work performed in chapters 4 and 5.

Figure 1.3 The relationships between the chapters comprising the thesis



Chapter 4 formulates a model of short text semantic similarity based on three factors: Propositional Content, Dialogue Act and Valence. It then goes on to develop classifiers for various types of question based on features drawn from function words. Consequently it supports the work in chapter 5 which optimises the Dialogue Act classifiers by clustering or grouping the function words by function (instead of alphabetically as in chapter 4). The discussion, conclusions and recommendations for future work in chapter 6 are supported by all of the preceding chapters.

Chapter 2 Production of STSS-65, the first Short Text Semantic Similarity dataset

For clarification, the production of the STSS-65 dataset and the corresponding section of (Li et al., 2006) is part of the programme of original work conducted solely by the author of this thesis. It is fully explained in (O'Shea et al., 2008) and (O'Shea et al., 2010).
STSS algorithms can play a crucial role in the development of Conversational Agents, Dialogue Systems in general and Information Retrieval.
As they are collections of multiple complex components, there is a need for benchmark datasets to evaluate STSS algorithms in isolation, before they are integrated into the complete system. An investigation at the onset of this work revealed no suitable benchmark datasets for performing such evaluation.
A literature search also showed no existing materials that were suitable for modification for this stage of the work. The closest materials were a set of medical diagnosis sentences (Rossell et al., 1988) and a set of sentences developed as part of a word similarity study (Miller and Charles, 1991). However, these sentence sets were not suitable for adoption as Rossell's set contained short sentences which were unlike real dialogue and inspection of the Miller & Charles set suggested there would be no high similarity pairs generated. Therefore an alternative approach was devised.
Consequently, this chapter describes the production of the first STSS dataset, consisting of 65 sentence pairs with human ratings, which is referred to as STSS-65. Following an analysis and critique of the present state of the art, the contributions fall into two categories: a methodology for producing STSS datasets and use of this methodology to produce STSS-65.

2.1 Contributions to research in the field

First, a procedure to create STSS datasets was designed combining the best known experimental techniques from prior work in word similarity. In the process a number of beliefs concerning word semantic similarity were examined. This chapter identifies those which were relevant for use in creating an STSS dataset.
Second, the procedure was used to create STSS-65, a robust dataset which has since been used in validating and benchmarking STSS algorithms. Since its publication (Li et al., 2006), it has been adopted as Gold Standard benchmark for STSS by a number of researchers (Kennedy and Szpakowitz, 2008, Islam and Inkpen, 2008 , Inkpen, 2007).

Third, the influence of varying the state of two important factors namely Order and Anchor (explained in section 2.8) on the collection of human ratings was investigated.
The data from these additional experiments was combined with data from STSS-65 in an Analysis of Variance (ANOVA). The outcome of the investigation was twofold. STSS-65 was validated, as it was found to produce good results which were very close to the best permutation of factors in the ANOVA analysis. However, the ANOVA analysis did discover one combination of factors which produced the highest precision amongst the human raters and this was statistically significant.
The final and possibly most important contribution of this chapter is to lay down a procedure by which other researchers can contribute to the extension of the set of STSS benchmarks. This is important for the exploration of additional features of the English language, such as the influence of Dialogue Acts on perceived similarity. Creation of additional STSS datasets is a highly labour intensive process and it is important to ensure that the new STSS datasets are compatible, for example so that they can be combined with integrity to produce larger training and testing partitions. The work in this chapter provides a clear definition of an objective, repeatable process to support this.
The intention is that the disappointing history of word pair similarity, where the only dataset which is widely accepted as credible is that of Rubenstein & Goodenough (Rubenstein and Goodenough, 1965), will not be repeated.

2.2 Obstacles to the creation of an STSS dataset

In the absence of an established benchmark STSS dataset, initial testing of STASIS was difficult and the need to produce such datasets was recognised. However, there are two fundamental obstacles to the creation of an STSS dataset. The first is selecting a sample that accurately represents the enormous range of Short Texts (STs) that can be composed from the English language. The second is to obtain numerical ratings that accurately represent human perception of similarity. Initially two primitive datasets were used, the Natural Language Processing (NLP) literature set and the Paired Dictionary (PD) set. The NLP literature set was developed by other members of the STASIS team, the PD set is an early stage of the work in this chapter. These will be discussed in the next two sections to illustrate the difficulties involved in creating a suitable STSS dataset.

2.2.1 The NLP Literature set

The NLP Literature set used a set of sentence pairs gleaned from NLP literature which were published as exemplars of varying degrees of similarity (Li et al., 2006). A table showing the sentence pairs with STASIS similarity ratings is reproduced in table 2.1.

Sentence Pair	Similarity	Sentence Pair	Similarity
1. I like that bachelor. I like that unmarried man.	0.561	2. I have a pen. Where do you live?	0
3. John is very nice. Is John very nice?	0.977	4. Red alcoholic drink. A bottle of wine.	0.585
5. It is a dog. That must be your dog.	0.739	6. Red alcoholic drink. Fresh orange juice.	0.611
7. It is a dog. It is a log.	0.623	8. Red alcoholic drink. An English dictionary.	0
9. It is a dog. It is a pig.	0.790	10. Dogs are animals. They are common pets.	0.738
11. I have a hammer. Take some nails.	0.508	12. Canis familiaris are animals. Dogs are common pets.	0.362
13. I have a pen. Where is ink?	0.129	14. Red alcoholic drink. Fresh apple juice.	0.420
15. A glass of cider. A full cup of apple juice.	0.678	16. I have a hammer. Take some apples.	0.121

Table 2.1 STASIS similarity measurements for the selected NLP sentence pairs

The fundamental problem of the NLP examples is that there was no set of human similarity measurements to compare the algorithm with. It relied on a qualitative judgment by the reader (Li et al., 2006) that "the similarities in the table are fairly consistent with human intuition."

A second problem is that NLP sentence pairs can be examples of difficult cases, chosen to stress test assumptions about the language. On inspection, the pairs in table 2.1 do not appear to be a representative test of the kind of short texts that algorithms may actually be called upon to compare in real dialogue.

2.2.2 The Paired Dictionary Set

The Paired Dictionary (PD) approach to producing an STSS dataset produced a more varied set of ST pairs by using a partly randomised sampling technique. However, if pairs of STs are sampled purely at random from the whole potential population then the likely outcome will be a Zipf-like power law distribution (Li, 1992) of similarity vs. frequency.

That is, there will be an extremely low proportion of high similarity pairs and a very high proportion of low similarity pairs. Unless a very large sample is taken, purely random sampling is unlikely to lead to any high similarity pairs at all.

To increase the likelihood of obtaining medium and high similarity pairs in the PD dataset, an ST pair was composed from dictionary definitions of the same word from two different dictionaries, the Concise Oxford English Dictionary (Pearsall, 2001) and the Collins English Dictionary (Hanks et al., 1986). One hundred target words for the dictionary lookup were selected using the book 'Alice's Adventures in Wonderland', by taking the first five occurring words beginning with 'a', 'b', 'c' and 'd' and the first four occurring words beginning with the other letters of the alphabet. This was with the exception of 'x' and 'z', which do not occur with a high enough frequency.

Words were only selected where a definition was of four or more words and adequately defined in a single phrase or sentence. Where multiple forms of the word were present, the first form encountered in the Oxford Dictionary was taken and the corresponding form found in the Collins.

The PD dataset attempted to overcome the fundamental problem of obtaining human ratings by using those from LSA, an established IR technique described in chapter 1. The assumption was that LSA would provide a ground truth similarity rating.

A sample of 10 of the definition pairs produced is shown in table 2.2. Two columns are used for the LSA ratings in the table. The Raw LSA column contains the ratings exactly as returned by the algorithm (-1 to +1), using the recommended topic space *General Reading up to 1st year college* and the maximum of 300 factors.

LSA is described a cosine measure which falls in the range -1 - +1 on the web interface for the LSA portal hosted by the University of Colorado at Boulder (the creators of LSA):

> "The system will compute a similarity score between -1 and 1 for each submitted sentence compared to next submitted sentence."

Because similarity ratings are usually reported on a range of 0 to +1, the Scaled LSA column reports the values scaled using equation 2.1:

$$ScaledLSA = \left(\frac{RawLSA + 1}{2} \right) \qquad (2.1)$$

This scaling is applied to the LSA ratings in succeeding chapters.

The ratings obtained showed large inconsistencies between STASIS and LSA. This meant that either STASIS, LSA or both were removed from the ground truth (human perception).

In fact, both of the techniques need to be tested for consistency with ground truth which can only be obtained from human ratings.

An additional problem of the PD set was the terseness of the definitions which was quite unnatural compared with real dialogue.

The NLP dataset was used to validate STASIS (Li et al., 2004). The PD set was not used in publications, but provided valuable experience. It focused attention on the need to produce a more realistic benchmark dataset using the best currently known experimental techniques to validate STSS measures.

Word	Definition	LSA Raw	LSA Scaled	STASIS
actually	as the truth of a situation as an actual fact	0.31	0.66	0.50
bank	An institution offering certain financial services such as the safekeeping of money. A financial establishment that uses money deposited by customers for investment.	0.62	0.81	0.43
conversation	An informal spoken exchange of news and ideas between two or more people. The interchange through speech of information.	0.21	0.61	0.61
day	A twenty-four hour period as a unit of time reckoned from one midnight to the next and corresponding to a rotation of the earth on its axis. The period of time the calendar day of twenty-four hours duration reckoned from one midnight to the next.	0.64	0.82	0.68
either	One or the other of two. Used before the first two alternatives specified the other being introduced by or.	0.43	0.72	0.32
feel	Perceive examine or search by touch. To perceive something by touching.	0.33	0.67	0.42
get	To come into possession of. Come to have or hold.	0.58	0.79	0.65
hot	Having a relatively high temperature. Having a high temperature.	0.99	1	0.80
into	To the interior or inner parts of. Expressing motion or direction to a point on or within.	0	0.5	0.41
just	Fair or impartial in action of judgment. Morally right and fair.	0.31	0.66	0.49
	Average	**0.44**	**0.72**	**0.53**

Table 2.2 A sample of STASIS and LSA similarity measurements for the PD dataset

2.3 Cutting the Gordian knot – the STSS-65 dataset

The difficulties in finding a method of validating an STSS measure were genuine and arose from the tangle of apparently intractable factors that emerge when attempting objective scientific measurement of attributes derived from human subjective experience. When measuring semantic similarity there is no known ground truth, consequently the only

option is to use the best experimental method (in terms of precision and accuracy) to obtain an overall human rating for pairs of items using a sample of the population. Secondly, the English language is capable of generating vast numbers and a great variety of short texts. Again it is necessary to obtain a representative sample, not only in terms of the distribution of semantic similarities but also in terms of the semantic and syntactic properties of the overall population of STs, if the results of a benchmark test are to generalise well. Therefore there are two major research questions which must be addressed in order to create an STSS dataset:

- How does one define a process of assigning a human rating to ST pairs (as a basis for comparison with an algorithm)?
- How does one select a representative sample of ST pairs from all the possibilities that can be generated by the English language?

The first question is addressed in this chapter and the second in following chapters.

2.4 Requirements of a Gold Standard dataset

The term "Gold Standard" originally applied to establishing the value of a currency and is associated with stability, transparency and reliability. It is also used to describe a testing method as being either the best possible or the best currently available. Datasets described as Gold Standard have been produced for related applications including question reformulation (Shaw et al., 2008), ontology mapping (Hu et al., 2008) and spoken dialogue summarisation (Gurevych and Strube, 2004). Strategies to achieve a genuine Gold Standard include training the raters (Hu et al., 2008), providing a coding manual (Wiebe et al., 1999) and allowing users to rate their level of confidence in their judgements (Su and Gulla, 2006). Emphasis is placed on agreement between multiple raters (Kilgarriff, 1998) and some studies may allow the human raters to negotiate a consensus rating (Su and Gulla, 2006). Alternatively, a bias correction mechanism has been applied to raters who disagree (Wiebe et al., 1999). Another strategy uses an iterative process in which an initial set of human ratings are scored against an existing Gold Standard with feedback before evaluations proceed (Kilgarriff, 1998). Finally, to reduce the labour intensive nature of creating materials for datasets, it is possible to use a combination of classifiers and find data items over which the classifiers disagree, then manually classify these before adding them to the Gold Standard dataset (Ngai and Yarowsky, 2000). Unfortunately, strategies

such as training, providing manuals and review by human experts are only applicable in domains where there is a known ground truth, which is not the case with semantic similarity. Also, high inter-rater agreement is difficult to achieve with more taxing combinations of data items. The requirements for a Gold Standard STSS dataset, drawing on prior work from word similarity, are discussed in the following sections.

2.4.1 Representative sampling

Any non-trivial benchmark dataset will be a sample of a larger overall population of data. As well as obtaining Short Texts (STs) that represent the overall semantic space of the English Language, it is important to obtain a sample of participants that represents the human population. Good representation of the general population is particularly pertinent for practical applications which could be used by anyone with web access. Neither of these issues has been given much consideration in earlier word studies.

The approach in this thesis is to take these two issues one step at a time. Regarding the first issue, because STSS-65 was highly novel, it was decided to use a simplified approach to collecting STs and accept known limitations to the data. The STs were taken from the dictionary definitions of the word pairs used by Rubenstein & Goodenough, to benefit from the large body of prior knowledge accumulated in the literature about the words. Regarding the second, STSS-65 uses a more representative sample (in terms of subject knowledge, age and gender) of the human population than previous word similarity studies.

2.4.2 Precision and Accuracy

STSS-65 consists of judgments by human participants. Precision requires the judgments to be in close agreement with each other. Accuracy requires the derived measures to be in close agreement with the "true" similarity. Precision is affected by both the participant's internal state (mental and physical) and the measurement instrument (for example ambiguity of instructions). Accuracy depends on a common human model of similarity and also on the possibility of blunders by the participant. These problems influence the design of the measurement instrument.

2.4.3 Measurement scale

The scale on which the similarity measures are made determines the statistical techniques that can be applied to them later (Blalock, 1979); the question is how sophisticated a measurement scale can be used?

It is a reasonable assumption that human intuitions of similarity are at least ordinal – that one pair of items can be more similar to each other than another pair. Interval scales improve on ordinal by having consistent units of measurement and ratio scales improve over interval by having an absolute zero point on the scale. STSS-65 is intended for algorithms that run from an absolute zero point (unrelated in meaning) to a maximum (identical in meaning) and thus is a form of ratio scale. Word semantic similarity has always been treated as a ratio scale attribute for both machine measures and human datasets. Setting the upper bound of the scale is common in word similarity measures and transformation of the range of the scale for comparisons is permissible.
The strength of measuring similarity on a ratio scale is the power of the statistical analyses that can be performed on ratio scale data such as correlation, Analysis of Variance (ANOVA) etc (Fenton and Pfleeger, 1998). The weakness is that the attribute of similarity is human and subjective; consequently it is difficult to measure experimentally as a ratio scale attribute.

2.5 The Rubenstein & Goodenough Legacy

The prior work which is most relevant to this thesis is the word semantic similarity study conducted in the mid-1960s by Rubenstein & Goodenough (Rubenstein and Goodenough, 1965). This experiment has been replicated on several occasions (Miller and Charles, 1991, Charles, 2000 , Resnik, 1999). These studies, coupled with that by Li et al (Li et al., 2006) provide evidence to support the view that human similarity measures are at least ordinal, showing reasonably consistent ranking between individuals, different groups of participants and over time. Collectively however, these replicated experiments have a number of uncontrolled factors which prevent them from being truly comparable. An analysis of these factors and their potential confounding effects forms the basis of the experimental programme conducted in this study.

2.5.1 Method of presentation of materials

Rubenstein & Goodenough printed each of the word pairs on a separate slip of paper and the subjects were asked to sort them in order of similarity before rating them. Miller & Charles (1991) presented all of their word pairs on two sheets of paper. Charles (2000) used a questionnaire with each word pair on a separate page. Resnik used an electronic version of the Miller & Charles (1991) questionnaire.

The different methods of presentation provide varying degrees of exposure to the dataset as a whole during the ratings of earlier items. It is possible that participants are uncertain about the process at the start and become "grounded" as they acquire a context from performing the ratings. If this is the case, the earlier items seen by each rater may be noisy until the rater has settled down. The advantage of seeing lots of pairs on a sheet (or screen) at the start is that grounding will be almost immediate reducing noise. However, if there are regular patterns in the presentation of blocks of data seen simultaneously, there is a greater risk of the earlier item pairs biasing later judgements. This can be mitigated by randomising the order of presentation.

The potential variation in the results between the early word similarity studies due to varied presentation of materials means that it is not possible to assume that results are comparable across the word similarity studies.

2.5.2 Method of randomisation of materials

Rubenstein & Goodenough shuffled the slips containing the word pairs into random order before presentation. Miller & Charles (1991) randomised the order of word pairs on the 2 sheets for each participant. Charles (2000) randomised the order of pages within his questionnaire for each participant. Resnik used two variants, a randomly selected ordering and a reversed version of that ordering.

None of the procedures describe randomisation of the order of words within a pair (e.g. coast-forest vs. forest-coast) despite prior work on asymmetry in similarity judgment (Tversky, 1977).

The advantage of thorough randomisation is that the likelihood of regular patterns in the order of presentation biasing the data will be lower. The disadvantage is that the process is more difficult to manage and vulnerable to operational error, particularly if paper questionnaires are used.

2.5.3 Instructions, guidance and measurement scale definition

Rubenstein & Goodenough instructed the participants to assign a value from 4.0 – 0.0 to each pair – the greater the similarity of meaning the higher the number. These instructions focus attention on the relative similarities of items in the dataset and may encourage expanding the range of similarity judgments to fill the range 4.0 to 0.0 even if other pairs with higher or lower similarity could exist outside of the dataset.

Miller & Charles (1991) and Resnik (1999) instructed the participants to examine each pair closely and then to rate it on a 5-point scale from 0 to 4, where 0 represents no similarity of meaning and 4 perfect synonymy. Charles (2000) advised participants to study each pair and to rate it for semantic similarity on a 5-point scale which ranged from similarity to dissimilarity. He also used semantic anchors for the 5 points ranging from 4: "identical in meaning" to 0: "opposite in meaning."

Charles was the only researcher to illustrate decreasing semantic similarity, using the following list of pairs: snake-serpent, snake–eel, snake-alligator, snake-frog, snake-book and snake-bulb. Resnik used the same instructions as Miller & Charles (1991).

Potential variation arises from encouraging the use of the first decimal place. Rubenstein & Goodenough clearly intended the use of the first decimal place, Miller & Charles (and Resnik) encourage the use of whole numbers (the first decimal place quoted in results may be a by-product of averaging).

Rubenstein & Goodenough used a general instruction about how to use the measurement scale whereas the others gave absolute descriptions of the end-points, allowing more variation in interpretation.

Rubenstein & Goodenough and Charles presented the word pairs for judgement one item at a time, whereas Miller & Charles and Resnik allowed the participants to see half of the word pairs as a complete block. This could encourage absolute assessment of similarity for the first two studies and relative assessment in the second two.

Finally, the use of semantic anchors in Charles (2000) could provide better interval measurement and lower noise than other methods.

Because the instructions varied, there could be corresponding variations in the results.

In fact there is no information published on the reactions of participants to these varying instructions. Consequently these variations may have introduced confounding factors making studies difficult to compare and begging the questions: "Which set of instructions is best in getting at the ground truth?" and "How much difference does the variation in instructions contribute to similarity ratings?"

2.5.4 Sampling the population for participants

In the work on word similarity that followed Rubenstein & Goodenough, no rationale or justification was given for choice of sample size; however all of them claimed statistical significance for at least one test in their results.

Rubenstein & Goodenough used two groups of college undergraduates for a total of 51 participants. Miller & Charles (1991) used 38 students; all specified as Native English speakers. Charles (2000) used two groups of undergraduates (50 participants on the 65 pair dataset and 58 on a 30 word pair subset). All were Native English Speakers and all received credit for courses taken in psychology. Resnik used 10 participants who were all computer scientists at graduate student or postdoctoral level. The lack of specificity in the published protocols prevents us from knowing:

- if Rubenstein & Goodenough and Resnik used only Native English speakers
- the academic background of students used by Miller & Charles and Resnik
- the academic level of the students used by Miller & Charles (1991)
- the gender or age composition of the groups.

These could be confounding factors in comparing experiments along with group size, although the review of word similarity literature suggests that only Resnik's sample size of 10 is likely to be a problem.

The most important potential issue is the high homogeneity of participants within the groups of particular experiments and their distinct differences from the general population which could reduce their value as representative samples.

2.5.5 Summary of work after Rubenstein & Goodenough

In every case, from the method of presenting the materials to the sampling of the population, the later researchers have introduced variations or interpretations of Rubenstein & Goodenough's methods. In some cases this is understandable as Rubenstein & Goodenough themselves did not publish a full account of what they did. Neither did they justify the choice of particular approaches. However, as section 2.5.3 and 2.5.4 illustrates, the later researchers are equally guilty. There are two consequences. The first is that none of the variations used emerges as the way which gets closest to ground truth. The second is that there is no strong justification for expecting results to be comparable across these experiments.

Part of the contribution of the work in this chapter is an investigation of the influence of these varying factors to determine how they should be fixed as experimental controls.

2.6 Creation of the STSS-65 dataset

Conducting an effective experiment to create a useful STSS dataset requires four fundamental steps. Materials must be collected, human ratings acquired, overall ratings calculated and the quality of the dataset validated. To achieve this, a number of choices must be made, including which materials (ST pairs) to acquire, how to construct a measurement instrument to administer to the participants and what statistical methods to use for measurement and evaluation. Constructing the measurement instrument requires selection of the measurement scale, composing the instructions to participants and procedural choices such as whether (and how) to randomise presentation of materials. All these important questions, which were addressed as part of a design process, are described in this section.

2.6.1 Selection of materials

The Rubenstein & Goodenough dataset (Rubenstein and Goodenough, 1965) contains 48 nouns arranged in various combinations to make up the 65 word pairs, for example *cord* appears twice, paired with both *smile* and *string*. The word pairs cover a range of word semantic similarity values from high (e.g. *cord-string*) to low (e.g. *cord-smile*). Although the Rubenstein & Goodenough dataset has an inherent bias to low word similarity, a subset of 30 word pairs (leaving out some low similarity pairs) was later selected which provided a representative coverage of the range of word similarities (Miller and Charles, 1991). Although no rationale has been published for the particular choices of words and pairings, the dataset is well known and is treated as a Gold Standard for comparing measures of word similarity.

The STSS-65 dataset builds on the knowledge gained by researchers from the Rubenstein & Goodenough word dataset, by adopting the dictionary definitions of the 65 word pairs as the materials and conducting the necessary experiments to obtain the similarity ratings. The Collins Cobuild Dictionary (Sinclair, 2001) was chosen over other dictionaries to provide the definitions for the new STSS-65 dataset because it is derived from the British National Corpus (400 million words at the time of compilation). Consequently the definitions take the form of natural sentences illustrating frequent usage. Other dictionaries

tend to make use of terse definitions. For example taking the word *string*, the Cobuild Dictionary's first definition is:

"String is thin rope made of twisted threads, used for tying things together or tying up parcels."

Whereas the Oxford English Dictionary's (Little et al., 1983) first definition is:

"A line, cord, thread."

Some Cobuild definitions span multiple sentences and for those cases the sentences were merged into a single one with subordinate clauses. For example, the definition of bird spans 3 sentences:

A bird is a creature with feathers and wings.
Female birds lay eggs.
Most birds can fly.

The merged version was:

A bird is a creature with feathers and wings, females lay eggs and most birds can fly.

The definition of smile was changed because it is circular:

A smile is the expression that you have on your face when you smile.

The alternative substitutes a fragment from the Cobuild definition of the verb smile into the original:

A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.

The definition of automobile:

An automobile is a car.

could not be modified because the definition of car was also a member of the dataset. Rooster-cock and Midday/Noon had genuinely identical definitions and were left unmodified.

The original Rubenstein & Goodenough word pairs, the derived sentence pairs, the word similarity scores for the word pairs and the sentence similarity scores are shown in table 2.4 which follows the description of the experimental procedure. There will also be a discussion of selected word pairs including those compiled from multiple sentences.

2.6.2 Similarity measurement scales

All of the word similarity experiments described previously used a Likert-like questionnaire-based process, asking the participants to write down a numerical similarity value for the pair of words on a scale from 0 to 4 (with procedural variations discussed in 2.5). Two alternative approaches were considered for this work before settling on the process:

- A questionnaire-based procedure using Thurstone scales
- A discriminability-based approach influenced by Signal Detection Theory

Likert-like and Thurstone approaches are well known in the Social Sciences as approaches for attitude measurement (Oppenheim, 1992). Signal Detection Theory is concerned with the classification of a signal being present or absent in the presence of noise (Egan, 1975) and has its origins in the training of radar operators during World War 2.
Thurstone Scales (Thurstone, 1927) and Signal Detection Theory (Egan, 1975) are measurement techniques arising from psychophysics which have gained general acceptance. The field of psychophysics involves investigation of methods to scale the judgements of individuals concerning a variety of physical properties, for example, weight (Oppenheim 1992).
Both are techniques founded on discriminable dispersion models, assuming that a group of humans will produce results following a normal distribution when asked to perform a rating or classification process; however, they have differing aims. Thurstone was concerned with generating scales which have equal-appearing intervals, whereas Signal Detection Theory (SDT) is concerned with finding the discriminability of a signal in the presence of noise, regardless of the criteria which might be used to make a classification by a human observer. Both have their merits in measuring STSS and both are labour-intensive for experimental participants.

2.6.2.1 Likert-like scales

Likert scales were developed to combine multiple responses from a participant to measure a single dimension, more specifically an attitude. An example of Likert scale construction illustrates the measurement of attitude towards children (Oppenheim, 1992). This requires the creation of an item pool, a set of statements indicating a particular feeling or position concerning children and asking the participants to score their own attitude towards the statements on a scale containing 5 points *strongly agree*, *agree*, *uncertain*, *disagree* and

strongly disagree. The attitudes are mapped onto a numeric scale typically from 5 (*strongly agree*) to 1 (*strongly disagree*) although the scoring may be reversed.
Scoring must be consistent with an overall positive or negative attitude towards children because the overall score is taken by summing the numerical responses to the questions. The items may mix statements with positive and negative positions on children. In this case a process known as "scoring reversal" must be used to ensure that (in this case) positive answers to negative items contribute negative components to the overall attitude. Scoring reversal is illustrated with two example items below (Oppenheim, 1992):

> *Children bring a husband and wife closer together*

and

> *Children need some of their natural meanness taken out of them*

These indicate (respectively) a positive and a negative attitude. If a high overall score indicates a positive attitude, a score of 5 assigned to the second item indicates a strong negative attitude which will need to be re-mapped to 1. Conversely a score of 1 will need to be re-mapped to 5, indicating a strong positive attitude to children.
Construction of the questionnaire requires collection and validation of the items. Some items may appear valid in the first instance, with difficulties becoming apparent during piloting. For example the item:

> *If we could afford to do so, we would prefer to send our children to boarding school*

may initially be seen as negative because the parents are avoiding the company of their children. However, it could also be explained by a parent's desire to give the child a good start in life (at some emotional sacrifice).
It may also be a poor item because it could confound the process by drawing in attitudes towards snobbery, social mobility and the social justice of private and exclusive education. Consequently Likert-scale items are purified and this can be achieved using a process similar to leave-one-out cross validation. Correlations are obtained between the scores for each individual item and those for the rest of the items. Those items which correlate poorly are discarded. The result is an attitude measurement instrument which is described as reliable and approaching unidimensionality in many cases (Oppenheim, 1992).

2.6.2.2 Thurstone scales

The purpose of constructing a measurement instrument using Thurstone scales is to achieve equal-appearing intervals between the points on the measurement scale. It is acknowledged (Oppenheim, 1992) that equal-appearing intervals may not be the same as a mathematically rigorous interval scale (such as the Celsius temperature scale). The technique, as described in (Oppenheim, 1992), is based on the ability to discriminate between items. An initial trial is conducted to discover "just noticeable differences" between the respondents, where a just noticeable difference corresponds to 50% of the participants being able to distinguish between two compared items. Consequently a difference of one unit on the scale, say between 3.0 and 4.0 means they lie one just noticeable difference apart in the psychological continuum being measured. Measuring pairwise differences between members of a large number of items results in a combinatorial explosion rendering experiments with human participants impractical. Thurstone's technique overcomes this problem using a less laborious sorting technique. The scale is produced by printing all of the items onto card and supplying a complete, randomised set to each judge (judges are expected to be a representative sample of the population the scale will be used with). The judges are presented with a hypothetical scale running between two extremes (e.g. most favourable to least favourable). The scale has a number of points (typically 11) and each point has a corresponding pile. The judges are then asked to sort cards into suitable piles (e.g. most favourable in pile 1 to least favourable in pile 11). The judges next number the items by pile and then, for each item, the complete set of judgements is analysed. Ideally, all the cards for a particular item would have the same rating, in practice there is a distribution of the ratings to be analysed. Items which have a broad distribution are decreed to be ambiguous and are discarded. Finally the medians are calculated for the (reasonably) pure items and from them a set is selected having medians which are nearest to the relevant scale points. This may be followed by a validation in which the statements are printed on a questionnaire in a randomised order and participants are asked to either agree or disagree with each statement. Again, ideally, a participant would agree with the single statement which most closely described the attitude being measured. In practice participants may agree with several. As long as these are few and consistent with a particular attitude the items can be accepted as valid (e.g. all agreed items mildly positive, disagreement with strongly positive, neutral and negative). However if participants agree with a particular item with one rating and a set of items with very different ratings it suggests that the particular item should be discarded.

It must be emphasised that all of the above work has to be performed **before** the scale can be used in a real experiment.

In use a participant selects the single attitude statement for each attribute being measured that most closely describes the participant's attitude to the attribute, to generate a numerical score. This is in contrast to the Likert scale, where degrees of agreement with members of a group of statements are combined to form the numerical score.

2.6.2.3 Selection of a scale technique

Each technique has its strengths and weaknesses; choosing one over the other entails a trade-off.

Likert scale experiments are more efficient than Thurstone scale experiments. Likert scales have been used to administer up to 350 items in one sitting whereas Thurstone scales have been used for up to 150 items (Oppenheim, 1992). Likert-like scales have been used in word similarity tasks with 10 - 38 judges (Resnik, 1999, Miller and Charles, 1991) in semantic similarity experiments, whereas Thurstone scales have used 40 -60 judges in sociological experiments (Oppenheim, 1992).

In terms of precision and accuracy, Thurstone scales are based on the approach of at least establishing equal-appearing scale points, an issue which is not addressed using Likert scales (Oppenheim, 1992). However, the difficulty of recruiting experimental participants and the number of experiments required counts heavily against the use of Thurstone scales in this work.

Blocked designs are easier with Likert-like scales, because each item is independent, whereas selection of the appropriate item on a Thurstone scale requires the participant to relate it to the rest of the items.

Both the traditional Likert and Thurstone approaches view the problem of rating semantic similarity from "the wrong end of the telescope." Effectively they use items to discover attitudes about the world, whereas similarity rating uses the attitudes of the participants, expressed in ratings, to discover information about the items.

A suitable modified Likert scale would consist of a single item of the form:

> *The two sentences are identical in meaning*

And the responses would be:

> *strongly agree*, *agree*, *uncertain*, *disagree* and *strongly disagree*

however, the **strength of the attitude** agreement/disagreement is not necessarily an indicator of the **degree** of similarity. Therefore, as a minimum, a modified description of the scale extremities would be required, with the new scale being described as a "Likert-like" scale.

In terms of statistical analysis the superficial judgement would be that Thurstone scales are superior. Likert scales per se only offer ordinal measurement whereas Thurstone scales appear to offer interval measurement. However, Thurstone scales may indeed offer better interval properties, but there is no guarantee that they will provide better mathematical rigour than Likert-like scales for which steps have been taken to improve the interval properties (discussed in section 2.5.3 and later in section 2.6.5).

The conclusion of this section is that the best approach to collecting ratings in the experiments in this section will be to begin with a Likert-like scale rating and investigate potential enhancements to improve the quality of the ratings to be collected. The Likert-like scale is modified so that instead of indicating strength of agreement with an attitude statement, the participant indicates judgement of the degree of an attribute (i.e. similarity). This is described fully in section 2.6.5.

2.6.3 Population and Sampling

The effort involved in conducting experiments with human participants is high, therefore a group size was required that would balance the need to provide statistically significant results with that of not using more participants than was necessary (this is also a requirement of the University ethics process).

The sample size chosen was based on the review of previous studies in section 2.5.4, with the proviso that it could be expanded if necessary.

A sample of 32 graduates was chosen to create the initial benchmark dataset in Experiment 1. This sample proved sufficient for statistically significant results (Li et al., 2004, O'Shea et al., 2008).

The aspiration was to represent the general population. However, because participants would be completing a 65-page (plus instructions) questionnaire in their own homes without supervision, it was decided to restrict the sample to people with graduate-level education. The sample was also restricted to Native English speakers, in common with many other verbal experiments in linguistics, psychology and neuroscience (Miller and Charles, 1991, Kemmerer, 2006, Bushell and Martin, 1997).

- Representation of the age range was good, considering the criterion that all must be educated to at least graduate level. The average age was 43.6 years and the standard deviation was 12.1 years (n=31, one participant withheld age).
- In terms of educational level, 14 held bachelor's degrees and 2 had equivalent professional qualifications. Five held master's degrees and 10 held PhDs. One participant withheld educational information (this does not cast doubt on meeting the requirement as all participants were recruited through contacts who knew their educational level).
- Portions of post-graduates vs. graduates were not important in this study. "Intelligence" was not a factor and the sole reason for using graduates was the belief that they would have the skills necessary to follow the instructions without supervision.
- Representation of both genders was achieved, the sample consisted of 21 males, 9 females and 2 withheld information.
- Representation of "the two cultures" (Snow, 1963) was achieved. This is important because the tradition of using students recruited from the researcher's department could lead to bias if the sample is skewed towards verbal (arts / humanities) or non-verbal (science / engineering) students. The graduates in this sample were composed of 18 from sciences, 9 from arts, 4 from disciplines straddling the boundary (e.g. Business Information technology) and 1 withheld information.

2.6.4 Randomisation

Each participant received the 65 sentence pairs in an individually randomised order. Also, sets of materials were produced in pairs (A and B), one of which had each pair of sentences in a particular order and the other of which had the reverse order (so A would have the sentence containing *coast* before the sentence containing *forest* and B would have *forest* before *coast*). Individual sentence pairs were then transferred between A and B at random to produce two new sets of materials containing mixtures of the A and B orders.

2.6.5 Procedures and Instructions

The participants were supplied with a package containing a questionnaire, instructions and an ethics statement. The package also contained materials for returning the questionnaire by post if personal contact was not convenient. Each package was arranged so that potential participants would read the ethics statement and decide whether or not to proceed

before accessing the questionnaire and instructions. Appendix 2 contains the following examples of experimental materials:

- Appendix 2.1 the ethics statement / covering letter
- Appendix 2.2 the instruction sheet
- Appendix 2.3 an individual question sheet (one of 65)
- Appendix 2.4 the personal data collection sheet.

The participants were presented with each sentence pair and asked to rate "how similar they are in meaning." The Likert-like scale for performing the rating was specified by defining the extreme ends of the measurement scale as "…between 0.0 (minimum similarity) and 4.0 (maximum similarity) on the form, please do not use values greater than 4.0." It was made explicitly clear on the front sheet that the first decimal place is meaningful with the statement "You can use the first decimal place (e.g. 2.2) to show finer degrees of similarity."

To improve the interval properties of the scale a technique similar to the Thurstone scale approach, developed by Charles (Charles, 2000), was used. Charles assigned a phrase described as a "semantic anchor", which was very similar to a Thurstone attitude statement. This allows the human rater to map an exact number to each of the 5 specific scale point descriptors or make an informed interpolation between two descriptors. Semantic Anchors for this work were chosen from Charles' set of 16 (Charles, 2000). This set expressed degrees of similarity and dissimilarity, running from *identical in meaning* to *opposite in meaning*. Charles' experiments scored the set of descriptors on a scale running from 0 (opposite) to 100 (identical).

The descriptors used in the experiments in this thesis were taken from the upper half of Charles' set; first by selecting the two anchors which best expressed the extremes of a scale measuring similarity only. Then 3 additional descriptors were selected to express the meanings of the intervening scale points. The best descriptors of the extremes were *unrelated in meaning* (scored at 44.3) and *identical in meaning* (scored at 100.0). The difference, 55.7, was divided by 4 to determine the desired intervals between the descriptors for a 5-point similarity scale yielding an interval of approximately 13.92. Following Thurstone's principle, the 5 semantic anchors which fell most closely to the major scale intervals were allocated. These are listed in table 2.3 along with the desired scores and the actual scores (from Charles' experiments) for the descriptors which became the semantic anchors used in this work.

Scale Point	Desired Score	Actual Score	Semantic Anchor
0.0	44.3	44.3	The sentences are unrelated in meaning.
1.0	58.22	58.0	The sentences are vaguely similar in meaning.
2.0	72.14	71.25	The sentences are very much alike in meaning.
3.0	86.08	88.1	The sentences are strongly related in meaning.
4.0	98.98 (100)	100	The sentences are identical in meaning.

Table 2.3 Semantic Anchors with desired and actual scale intervals

2.6.6 Results

Table 2.4 shows the results of experiment 1. The first two columns contain the original word pairs and similarity ratings from the Rubenstein & Goodenough experiment. The second pair of columns contains the sentence pairs created for this study with the similarity ratings obtained. In both cases the similarity ratings are on a scale from 0 – 4.0.

Word pair	R&G Word Sim	Sentence pair	STSS-65 Sentence Sim
1. cord **smile**	0.02	Cord is strong, thick string. A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.	0.04
2. rooster voyage	0.04	A rooster is an adult male chicken. A voyage is a long journey on a ship or in a spacecraft.	0.02
3. noon string	0.04	Noon is 12 o'clock in the middle of the day. String is thin rope made of twisted threads, used for tying things together or tying up parcels.	0.05
4. fruit furnace	0.05	Fruit or a fruit is something which grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat. A furnace is a container or enclosed space in which a very hot fire is made, for example to melt metal, burn rubbish or produce steam.	0.19
5. autograph: **shore**	0.06	An autograph is the signature of someone famous which is specially written for a fan to keep. The shores or shore of a sea, lake or wide river is the land along the edge of it.	0.02
6. automobile wizard	0.11	An automobile is a car. In legends and fairy stories, a wizard is a man who has magic powers.	0.08
7. mound stove	0.14	A mound of something is a large rounded pile of it. A stove is a piece of equipment which provides heat, either for cooking or for heating a room.	0.02

8. grin implement	0.18	A grin is a broad smile. An implement is a tool or other piece of equipment	0.02
9. asylum **fruit**	0.19	An Asylum is a psychiatric hospital. Fruit or a fruit is something which grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat.	0.02
10. asylum monk	0.39	An Asylum is a psychiatric hospital. A monk is a member of a male religious community that is usually separated from the outside world	0.15
11. graveyard madhouse	0.42	A graveyard is an area of land, sometimes near a church, where dead people are buried. If you describe a place or situation as a madhouse you mean that it is full of confusion and noise.	0.09
12. glass magician	0.44	Glass is a hard transparent substance that is used to make things such as windows and bottles. A magician is a person who entertains people by doing magic tricks	0.03
13. boy **rooster**	0.44	A boy is a child who will grow up to be a man. A rooster is an adult male chicken.	0.43
14. cushion jewel	0.45	A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable. A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.	0.21
15. monk slave	0.57	A monk is a member of a male religious community that is usually separated from the outside world. A slave is someone who is the property of another person and has to work for that person.	0.18
16. asylum cemetery	0.79	An Asylum is a psychiatric hospital. A cemetery is a place where dead people's bodies or their ashes are buried.	0.15
17. coast forest	0.85	The coast is an area of land that is next to the sea. A forest is a large area where trees grow close together.	0.19

18. grin lad	0.88	A grin is a broad smile. A lad is a young man or boy.	0.05
19. shore woodland	0.9	The shores or shore of a sea, lake or wide river is the land along the edge of it. Woodland is land with a lot of trees.	0.33
20. monk oracle	0.91	A monk is a member of a male religious community that is usually separated from the outside world. In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth.	0.45
21. boy **sage**	0.96	A boy is a child who will grow up to be a man. A sage is a person who is regarded as being very wise.	0.17
22. automobile cushion	0.97	An automobile is a car. A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable	0.08
23. mound shore	0.97	A mound of something is a large rounded pile of it. The shores or shore of a sea, lake or wide river is the land along the edge of it.	0.14
24. lad wizard	0.99	A lad is a young man or boy. In legends and fairy stories, a wizard is a man who has magic powers.	0.13
25. forest **graveyard**	1.0	A forest is a large area where trees grow close together. A graveyard is an area of land, sometimes near a church, where dead people are buried.	0.26
26. food **rooster**	1.09	Food is what people and animals eat. A rooster is an adult male chicken.	0.22
27. cemetery **woodland**	1.18	A cemetery is a place where dead people's bodies or their ashes are buried. Woodland is land with a lot of trees.	0.15

28. shore voyage	1.22	The shores or shore of a sea, lake or wide river is the land along the edge of it. A voyage is a long journey on a ship or in a spacecraft.	0.08
29. bird **woodland**	1.24	A bird is a creature with feathers and wings, females lay eggs and most birds can fly. Woodland is land with a lot of trees.	0.05
30. coast hill	1.26	The coast is an area of land that is next to the sea. A hill is an area of land that is higher than the land that surrounds it.	0.40
31. furnace implement	1.37	A furnace is a container or enclosed space in which a very hot fire is made, for example to melt metal, burn rubbish or produce steam. An implement is a tool or other piece of equipment.	0.20
32. crane rooster	1.41	A crane is a large machine that moves heavy things by lifting them in the air. A rooster is an adult male chicken.	0.08
33. hill **woodland**	1.48	A hill is an area of land that is higher than the land that surrounds it. Woodland is land with a lot of trees.	0.58
34. car journey	1.55	A car is a motor vehicle with room for a small number of passengers. When you make a journey, you travel from one place to another.	0.29
35. cemetery mound	1.69	A cemetery is a place where dead people's bodies or their ashes are buried. A mound of something is a large rounded pile of it.	0.23
36. glass jewel	1.78	Glass is a hard transparent substance that is used to make things such as windows and bottles. A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.	0.43
37. magician **oracle**	1.82	A magician is a person who entertains people by doing magic tricks. In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth.	0.52

38. crane implement	2.37	A crane is a large machine that moves heavy things by lifting them in the air. An implement is a tool or other piece of equipment.	0.74
39. brother lad	2.41	Your brother is a boy or a man who has the same parents as you. A lad is a young man or boy.	0.51
40. sage wizard	2.46	A sage is a person who is regarded as being very wise. In legends and fairy stories, a wizard is a man who has magic powers.	0.61
41. oracle **sage**	2.61	In ancient times, an oracle was a priest or priestess who made statements about future events or about the truth. A sage is a person who is regarded as being very wise.	1.13
42. bird crane	2.63	A bird is a creature with feathers and wings, females lay eggs and most birds can fly. A crane is a large machine that moves heavy things by lifting them in the air.	0.14
43. bird cock	2.63	A bird is a creature with feathers and wings, females lay eggs and most birds can fly. A cock is an adult male chicken.	0.65
44. food fruit	2.69	Food is what people and animals eat. Fruit or a fruit is something which grows on a tree or bush and which contains seeds or a stone covered by a substance that you can eat.	0.97
45. brother monk	2.74	Your brother is a boy or a man who has the same parents as you. A monk is a member of a male religious community that is usually separated from the outside world.	0.18

46. asylum madhouse	3.04	An Asylum is a psychiatric hospital. If you describe a place or situation as a madhouse you mean that it is full of confusion and noise.	0.86
47. furnace **stove**	3.11	A furnace is a container or enclosed space in which a very hot fire is made, for example to melt metal, burn rubbish or produce steam. A stove is a piece of equipment which provides heat, either for cooking or for heating a room.	1.39
48. magician **wizard**	3.21	A magician is a person who entertains people by doing magic tricks. In legends and fairy stories, a wizard is a man who has magic powers.	1.42
49. hill **mound**	3.29	A hill is an area of land that is higher than the land that surrounds it. A mound of something is a large rounded pile of it.	1.17
50. cord **string**	3.41	Cord is strong, thick string. String is thin rope made of twisted threads, used for tying things together or tying up parcels.	1.88
51. glass **tumbler**	3.45	Glass is a hard transparent substance that is used to make things such as windows and bottles. A tumbler is a drinking glass with straight sides.	0.55
52. grin **smile**	3.46	A grin is a broad smile. A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.	1.94
53. serf **slave**	3.46	In former times, serfs were a class of people who had to work on a particular person's land and could not leave without that person's permission. A slave is someone who is the property of another person and has to work for that person.	1.93

54. journey **voyage**	3.58	When you make a journey, you travel from one place to another. A voyage is a long journey on a ship or in a spacecraft.	1.44
55. autograph **signature**	3.59	An autograph is the signature of someone famous which is specially written for a fan to keep. Your signature is your name, written in your own characteristic way, often at the end of a document to indicate that you wrote the document or that you agree with what it says.	1.62
56. coast **shore**	3.60	The coast is an area of land that is next to the sea. The shores or shore of a sea, lake or wide river is the land along the edge of it.	2.35
57. forest **woodland**	3.65	A forest is a large area where trees grow close together. Woodland is land with a lot of trees.	2.51
58. implement **tool**	3.66	An implement is a tool or other piece of equipment. A tool is any instrument or simple piece of equipment that you hold in your hands and use to do a particular kind of work.	2.36
59. cock **rooster**	3.68	A cock is an adult male chicken. A rooster is an adult male chicken.	3.45
60. boy **lad**	3.82	A boy is a child who will grow up to be a man. A lad is a young man or boy.	2.32
61. cushion **pillow**	3.84	A cushion is a fabric case filled with soft material, which you put on a seat to make it more comfortable. A pillow is a rectangular cushion which you rest your head on when you are in bed.	2.09

62. cemetery **graveyard**	3.88	A cemetery is a place where dead people's bodies or their ashes are buried. A graveyard is an area of land, sometimes near a church, where dead people are buried.	3.09
63. automobile **car**	3.92	An automobile is a car. A car is a motor vehicle with room for a small number of passengers.	2.23
64. midday **noon**	3.94	Midday is 12 o'clock in the middle of the day. Noon is 12 o'clock in the middle of the day.	3.82
65. gem **jewel**	3.94	A gem is a jewel or stone that is used in jewellery. A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.	2.61

Table 2.4 The STSS-65 dataset with corresponding word pair ratings from Rubenstein & Goodenough

2.7 Discussion of the STSS-65 dataset

2.7.1 Overall consistency between STSS-65 and Rubenstein & Goodenough

The motivation for using the Rubenstein & Goodenough word similarity dataset was to benefit from the decades of experience with that dataset. The underlying conjecture was that if two words have a particular degree of word similarity, their definitions should have a consistent degree of sentence similarity. So the question that arises is "How consistent are the word and sentence pair similarities over the two datasets?"

An appropriate measure of consistency is the correlation coefficient. Previous consideration of measurement scale properties showed that similarity measurements are usually treated as being on a ratio scale (Rubenstein and Goodenough, 1965, Miller and Charles, 1991, Charles, 2000 , Resnik, 1999) and for this the Pearson Product-Moment correlation coefficient is suitable(Blalock, 1979). However, it was also clear that similarity measurements can not achieve the purity of concrete physical measures such as temperature and electrical current. Therefore it is also worth taking a second opinion, in the form of Spearman's ρ (Fenton and Pfleeger, 1998), the Rank correlation coefficient for ordinal data.

Calculating these for the 65 pairs of similarity ratings (words vs. sentences) the results are:

- Pearson's r = 0.847 (P-Value = 0.000)
- Spearman's ρ = 0.887

For both r and ρ, a value of +1 indicates perfect correlation, 0 indicates no relationship and -1 indicates a perfect negative correlation. P-values indicate the likelihood of obtaining the result by chance.

The sentence dataset created for this work was observed to have a similar bias towards low frequency item pairs to Rubenstein & Goodenough's original word set (Miller and Charles, 1991). Moreover, examination of the low similarity portion of the data shows a lot of small variations between items, without any discernible pattern. Inspection showed that, for all practical purposes, these small variations are simply noise.

A similar approach to that of Miller & Charles (Miller and Charles, 1991) (for word pairs) was taken when applying it to evaluate sentence similarity measures. A 30 sentence-pair subset was extracted to span the range of similarities more evenly. The approach taken was to sample at equal intervals (one in every 4 sentence pairs) across the low end of the similarity range and take all of the ratings from the 47th to the 65th. This does not result in a

perfectly even distribution but corrects the bias to a reasonable degree. The sampled pairs are shown in bold type in table 2.4. The results for the subset of 30 pairs are:

- Pearson's r = 0.855 (P-Value = 0.000)
- Spearman's ρ = 0.944

So when the effect of noise in the low similarity pairs has been reduced, there is a significant correlation between the similarities for the word pairs and the sentence pairs. Also there is strong agreement on the ordering of the similarity ratings.
It could be asked why the two sets of ratings are not in perfect agreement. Before seeking explanations, it should be considered that it is unreasonable to expect perfect agreement (correlation = 1.0). Even when the Rubenstein & Goodenough word experiments were replicated (using the Miller & Charles 30-word subset) correlations of 0.97 (Miller and Charles, 1991) and 0.96 (Resnik, 1999) were obtained.
For some ratings there is a large difference between the word pair similarity and the sentence similarity, and these may be explained by polysemy. Examples of this include glass-tumbler, hill-mound, journey-voyage, autograph-signature, magician-wizard, cushion-pillow and furnace-stove. It has been suggested (Tversky, 1977) that when asked to compare two words for similarity, the participants access the various word senses in their lexical memory and select those that are closest in meaning. Glass-tumbler and magician-wizard are good examples of this. The definitions selected for polysemous words are those of the most frequent senses, which are given first in the list of definitions in the Cobuild dictionary. So the definition of tumbler is:

A tumbler is a drinking glass with straight sides.

But the definition of glass is the definition of the substance glass and it explicitly defines two uses which do not include the use as a drinking vessel:

Glass is a hard transparent substance that is used to make things such as windows and bottles.

However, according to Tversky's proposal (Tversky, 1977), it is possible that the human raters in the Rubenstein and Goodenough experiment are accessing the polysemous sense:

A glass is a container made from glass, which you can drink from and which does not have a handle.

It is also likely that they have blocked the even less similar polysemous sense of tumbler:

A tumbler is an acrobat, who performs on the ground often with other members of a group.

The definition of wizard is:

In legends and fairy stories, a wizard is a man who has magic powers.

But the definition of magician is much more mundane:

A magician is a person who entertains people by doing magic tricks.

However, it is again possible that the human raters in the Rubenstein and Goodenough experiment are accessing the polysemous sense:

In fairy stories, a magician is a person, usually a man, who has magic powers.

Additionally, both *wizard* and *magician* share a lower frequency sense in which the word is used to describe admiration for someone's skill or ability. So again it is possible that the human raters in the Rubenstein & Goodenough word experiment have selected the sense for each word that returns the highest similarity.

All of the remaining ratings with the high discrepancy between the word pairs and sentence pairs appear to have scope for this disambiguation factor, to a lesser degree (for example the Oxford dictionary contains a minor definition of *mound* which includes the word "hillock"). Another interesting observation is that like automobile-car, three of them include one of the words in the definition of the other and have relatively high standard deviations. These are journey-voyage (SD 1.09), autograph-signature (SD 1.29) and cushion-pillow (SD 1.06).

2.7.2 The challenge of medium similarity

If all humans had the same, perfect, concept of semantic similarity one might expect all the human participants to give exactly the same numerical rating to each sentence pair. Given human nature, a more reasonable expectation is for a distribution approximating to a normal distribution with the proviso that placing bounds on the range of similarities will lead to skewed or truncated distributions.

Figure 2.1 shows a histogram for sentence pair 36, which has a majority agreement on a very low similarity, tailing off very rapidly. Figure 2.2 shows a histogram for sentence pair 59, which has a majority agreement on a very high similarity, with a slightly less regular tail reaching into the medium similarity zone.

Figure 2.1 Histogram of similarity ratings for sentence pair 36

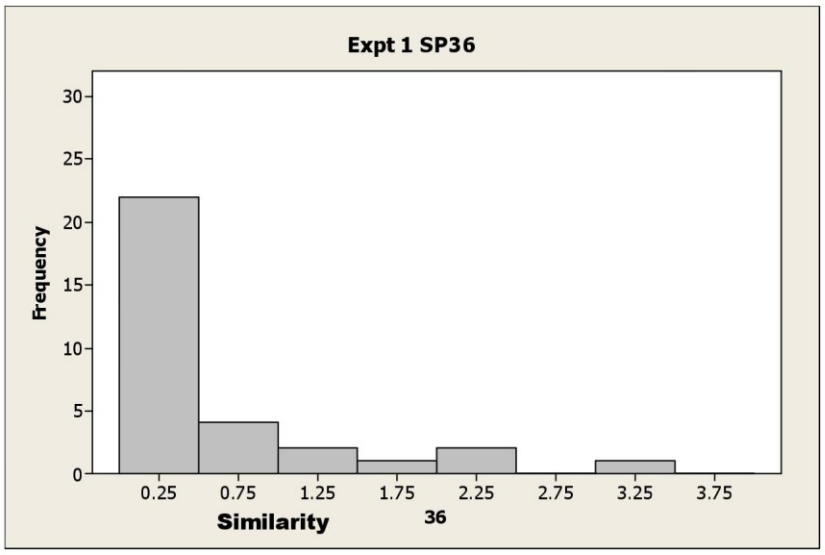
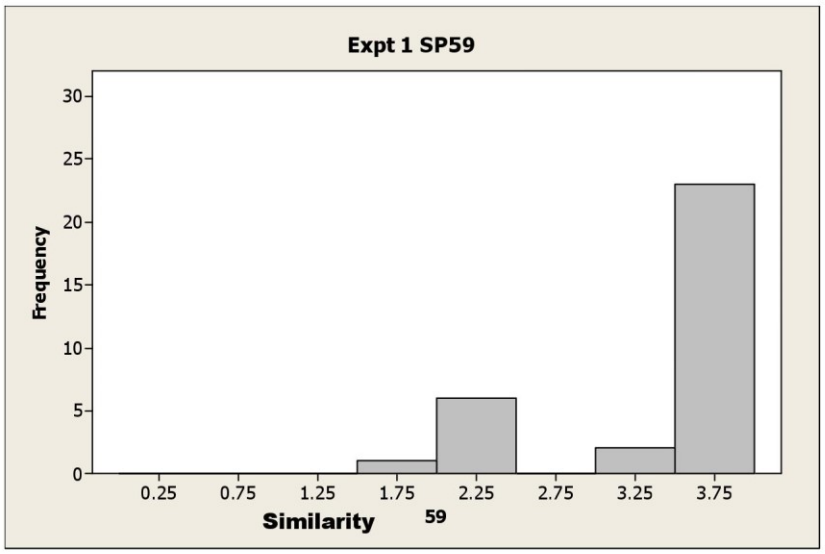


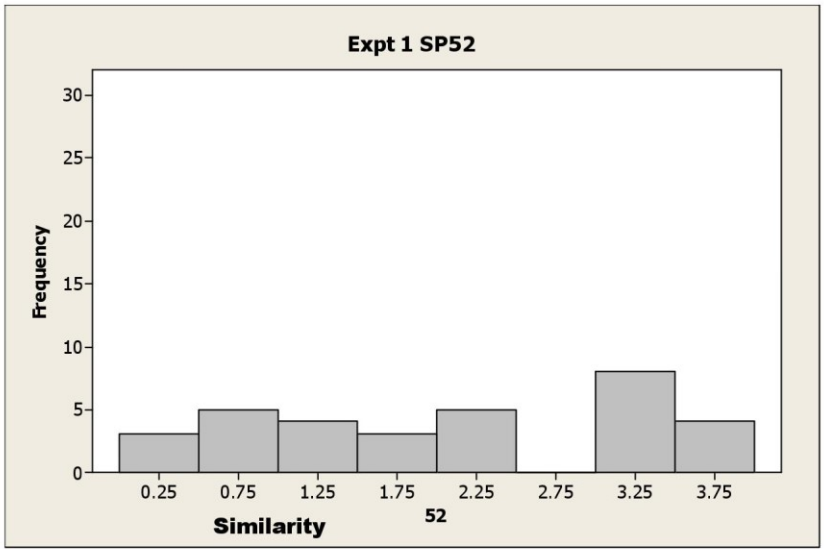
Figure 2.2 Histogram of similarity ratings for sentence pair 59



Both of the above cases are consistent with the idea of ratings following a normal distribution. With increased numbers of ratings, each could tend to half of a normal distribution cut off at either the low similarity or high similarity ends of the measurement range.

Therefore, one would reasonably expect medium similarity sentence pairs to have graphs approximating a normal distribution with the mean close to a peak occurring in the medium similarity zone. However, medium similarity sentence pairs seem to be generally characterized by an inconclusive set of judgments spread across the similarity range, the almost flat distribution of sentence pair 52 (figure 2.3) being an extreme example of this.

Figure 2.3 Histogram of similarity ratings for sentence pair 52



Medium similarity is used as a concept and referred to in a number of studies in various fields. However, there seems to have been no investigation of the meaning of medium similarity, its implications and applications to date. Empirically, Rubenstein & Goodenough commented briefly that participants appeared to react differently to the medium similarity part of their measurement scale (compared to high similarity) but they also noted the reaction was consistent between two different experimental groups (Rubenstein and Goodenough, 1965). This is a very interesting area for future investigation; however it does not form part of the essential focus of the present work. The evidence from this section supports the position that the sentence similarity ratings are credible when background knowledge about word similarities is taken into account and there are plausible explanations for any discrepancies which do not undermine the sentence similarity ratings.

2.8 Validation of the STSS-65 dataset

The primary aim of this section is the investigation of the influence of two factors, Order and Anchor, which were subject to variation across previous word similarity studies, on the outcomes of experiments measuring STSS. The contribution of this section is twofold. First, the outcome of the experiments that were conducted provides good evidence to validate STSS-65 and justify its adoption as a Gold Standard. Second, it provides evidence for modifications to optimise the experimental procedure leading to improvements in precision and accuracy.

Order is concerned with (the randomisation of) the order of presentation of sentence pairs and Anchor is concerned with the information supplied to improve the interval properties of the ratings obtained.

The two different states of Order investigate the difference between the sorting approach of the Rubenstein & Goodenough experiment and the questionnaire form used in the replications. The two different states of Anchor investigate the difference between Charles (2000) in using semantic anchors to describe the major similarity scale intervals and the other experiments which did not.

Given the possibility of an interaction between Order and Anchor, it was decided to conduct a 2-level, 2-factor ANOVA analysis. Four experiments (experiments 2.1-2.4) were required to collect data for each permutation of the factors. Experiment 2.1 simply involved extracting the data for the first 18 participants from experiment 1 (reported in Li et al.) (Li et al., 2006); experiments 2.2, 2.3 and 2.4 involved the collection of new data. Because the participants for a given level of a factor are split across two groups, there are actually 36 participants for each level, so the technique is more powerful than the group size of 18 suggests.

In order to study these factors correctly, a number of potentially confounding factors were treated as control variables. In particular these were randomisation, instructions and participants.

Conducting an ANOVA analysis requires certain conditions to be satisfied for the test to be valid. As the conditions can only be tested after the data has been collected, these are covered in section 2.9.3.

Appendix 2 contains copies of the variants of the experimental materials required for experiments 2.2, 2.3 and 2.4. These are:

- Appendix 2.5 modified instructions for card sorting experiments 2.2 and 2.3
- Appendix 2.6 a sample card
- Appendix 2.7 sample rating recording sheet for experiment 2.3 (2.2 omits the semantic anchors)
- Appendix 2.8 a modified instruction sheet for experiment 2.4 (semantic anchors removed).

2.8.1 Common Instructions and variations

Some variation in instructions was required as two experiments involved sorting cards and two involved questionnaires. In all experiments the task was described as "rate the similarity of meaning." Participants were asked to do this by "writing a number between

0.0 (minimum similarity) and 4.0 (maximum similarity) on the form, please do not use values greater than 4.0. You can use the first decimal place (e.g. 2.2) to show finer degrees of similarity."

For the questionnaire forms participants were instructed to work through from start to end without going back to revise earlier judgments. For the card sort participants were asked to

"…start by reading through the cards in the order you got them in, thinking about the similarity of the meanings of the two sentences on each card. Now please sort the cards in a rough order of the similarity of meaning of the sentence pairs"

before making a second pass through the cards to record their judgments.

The method of randomisation, of the order of the pairs of sentences and the order of the sentences within a pair, was the same as that described in section 2.6.4. Whenever a pack of cards was re-used, it was checked to ensure that no marks had been made by participants, and then shuffled to randomise it again, before issuing it to a new participant.

2.8.2 Population and Sampling

The aspiration was to represent the general population. However the participants would be performing the task without supervision and as in experiment 1, it was decided to restrict the sample to people who were Native English speakers with graduate-level education. Overall gender balance was achieved across the study with 51% males, 41% females and 8% withholding information. The balance between males and females was generally achieved within the experimental group samples.

2.8.3 Age

The overall average age was 41.9 (SD = 10.1) years (5 participants withheld their age). The range of average ages for the experimental groups was 39.7 (SD = 8.7) – 43.6 (SD = 12.8) years. Thus the age distribution was more diverse than in prior word studies.

2.8.4 Education

All but one participant (who had an equivalent professional qualification) had a bachelor's degree. The overall breakdown by academic background was 50% Science & Engineering, 39% Arts & Humanities, 8% mixed Arts/Sciences (e.g. Architecture) and 3% withheld information. The overall breakdown of qualifications was of 56% bachelor's degrees, 21% PhDs, 15% Masters, 4% PGCE with the remainder being either professional (graduate equivalent) qualifications or withheld.

In 3 groups the balance was approximately two-thirds bachelor level and one-third postgraduate degree. In the remaining group the balance was approximately reversed. The distribution between graduates and postgraduates is not considered important as discussed in section 2.6.3. In 3 groups there was an almost perfect balance between arts and sciences backgrounds, the remaining group was moderately biased towards sciences.

2.9 Experimental factors

2.9.1 Order

The Order factor had two levels, *quest* and *card*. The *card* variant used a deck of 65 stiff cards, with one sentence pair printed on each. The *quest* variant presented the sentence pairs on a questionnaire, with one pair on each sheet. Judgments were recorded on the sheet before moving on. In all experiments the question pairs were identified by code strings for ease of transcription; their values were allocated so as not to imply any pre-conceived similarity values.

2.9.2 Anchor

The Anchor factor had two levels, w*ith* and *without* semantic anchors. The *with* variant describes the scale points with 5 semantic anchors listed in section 2.6.5, table 2.3.
The structure of the four experiments making up this factor-based study is shown in table 2.5.

Experiments	With Anchors	Without Anchors
Questionnaire	Expt. 2.1 Questionnaires with Semantic Anchors	Expt. 2.2 Questionnaires without Semantic Anchors
Card Sort	Expt 2.4 Card Sort with Semantic Anchors	Expt 2.3 Card Sort without Semantic Anchors

Table 2.5 The experiments comprising the study

2.9.3 Requirements for valid ANOVA analysis

The two requirements for the results of an ANOVA analysis to be meaningful are:

- The data should follow a normal distribution

- The datasets collected for the different combinations of factors should have equal variances.

2.9.3.1 Test for normal distribution

Testing for a normal distribution can be performed visually, using a histogram and a probability plot. A problem with producing meaningful graphs was that, potentially, each sentence pair in the data has a different mean and that for each sentence pair there are relatively few ratings. Therefore it was decided to:

- Use the subsample of 30 sentence pairs to avoid bias to low similarity responses
- Use 39 ratings for each sentence pair (by including questionnaires that were returned after the deadline for the key publication date for STSS-65)
- Normalise the ratings for the sentence pairs so that each sentence pair had a mean of 0
- Combine the normalised ratings for all of the sentence pairs.

Consequently it was possible to produce graphs using a set of 1,170 ratings. However, there is a further consideration.

This data is constrained to the range 0.0 to 4.0. Assuming the original responses are drawn from a normal distribution, it is reasonable to expect a symmetrical response about the mean for sentence pairs in the middle of the range (similarity $\approx$ 2.0). However, ratings close to the extremes of the range (> 3.0, <1.0) will have one side of the distribution truncated because it simply can not go above 4.0 or below 0.0. Therefore, even if the samples are coming from a population which is naturally normally distributed, a degree of Kurtosis (peakiness) is to be expected. This is exactly what happened in the histogram plot shown in figure 2.4.

The second test is to produce a probability plot. This was conducted with Minitab and the plot is shown in figure 2.5. The p-value for the probability plot is not consistent with a normal distribution, but the plot does pass the well-known "fat pencil" rule of thumb used by engineers to assess normality (the "fat pencil" test is so called because if all the points in the probability plot fall around a straight line which could be covered by a fat pencil the distribution is taken as normal).

Two more factors concerning normality should be considered. First, the ANOVA technique is considered to be robust by statisticians, that is it can tolerate moderate departures from normality and homogeneity of variance (Blalock, 1979). Second, the

ratings arising from the experiments are themselves means of sets of human ratings, therefore the overall set of ratings is likely to tend to a normal distribution in accordance with the central limit theorem (Rice, 1994).

Figure 2.4 Histogram of normalised sentence pair ratings

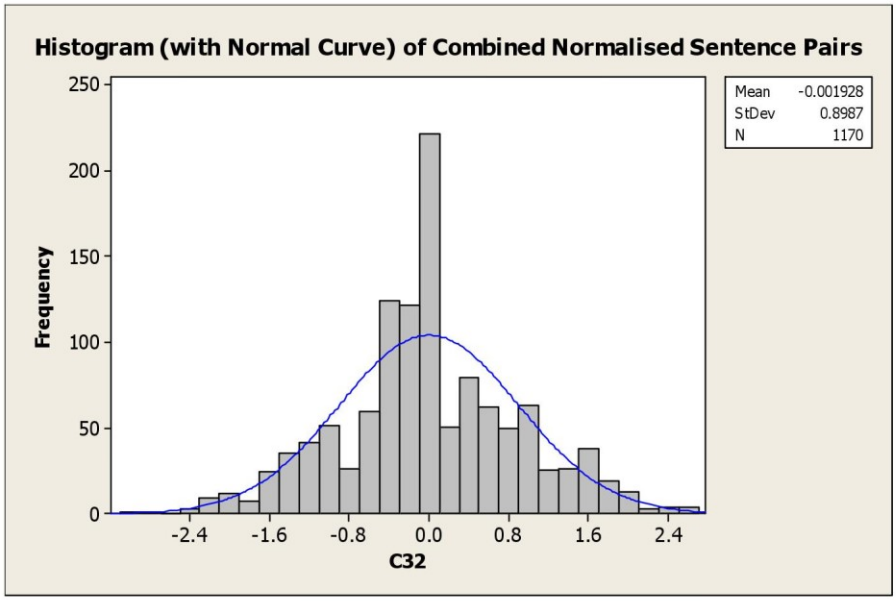
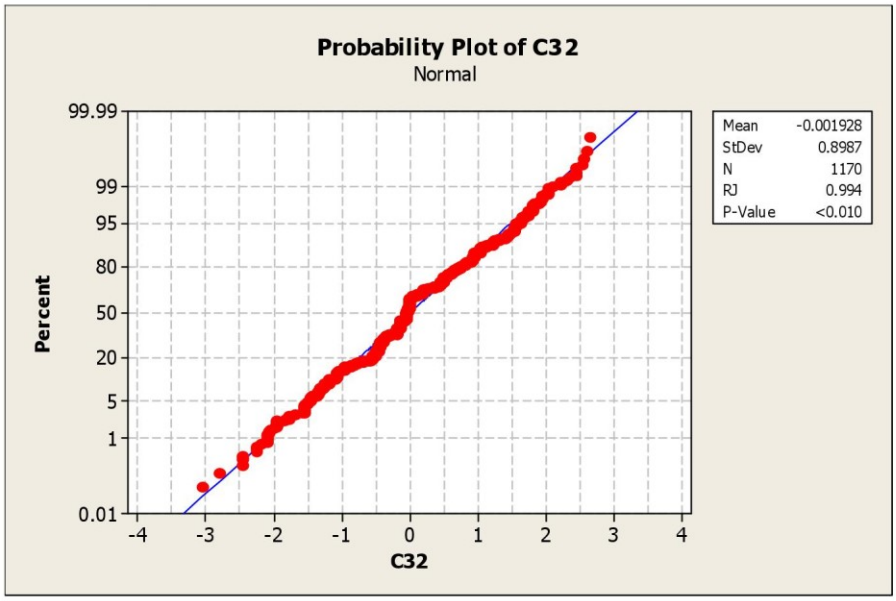


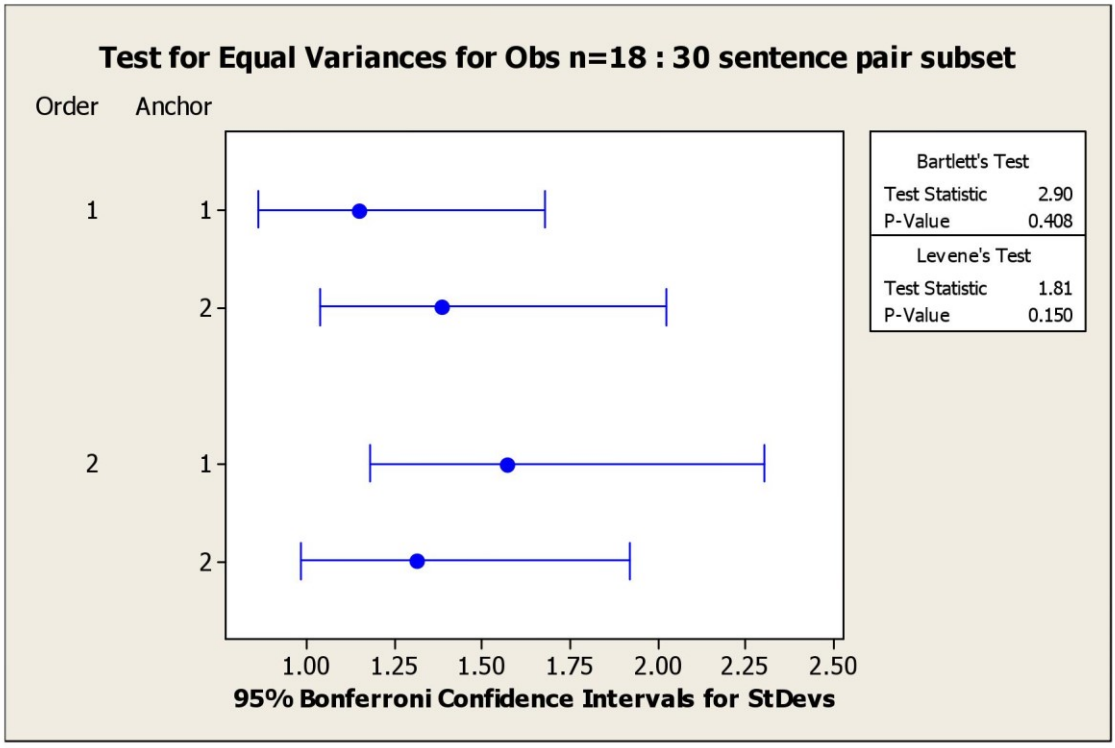
Figure 2.5 Probability Plot of normalised sentence pair ratings



2.9.3.2 Test for equal variances

Two tests were applied for homogeneity of variances, Levene's (Levene, 1960) and Bartlett's (Snedecor and Cochran, 1989). Results for the tests are displayed along with a confidence interval plot by Minitab, as shown in figure 2.6.

Figure 2.6 Results for Bartlett's test and Levene's test for equal variances



Examining homogeneity of variances, the p-values are Bartlett's = 0.408 and Levene's = 0.150. In both cases H_0 is an assumption of homogeneous variances and there is insufficient evidence to reject this. As a final note on sensitivity, the Minitab online documentation states that "The ANOVA F-test is only slightly affected by inequality of variance if the model contains fixed factors only and has equal or nearly equal sample sizes." This model has fixed factors and exactly equal sample sizes, so it is reasonable to assume equal variances and perform the ANOVA analysis.

2.9.4 ANOVA Analysis of Order and Anchor

The 2-level, 2-factor design was implemented as a 1-factor, 4-level General Linear Model in Minitab, combining all 65 sentence pairs in a blocked design to analyse the results. A total of 72 participants was distributed into 4 groups of 18, one for each combination of factors. The results were Order F = 7.49, P = 0.007 and Anchor F = 63.67, P= 0.000 indicating both Order and Anchor factors were significant. However, Order.Anchor was reported as F= 3.22, P = 0.074 indicating that the interaction between Order and Anchor was not significant.

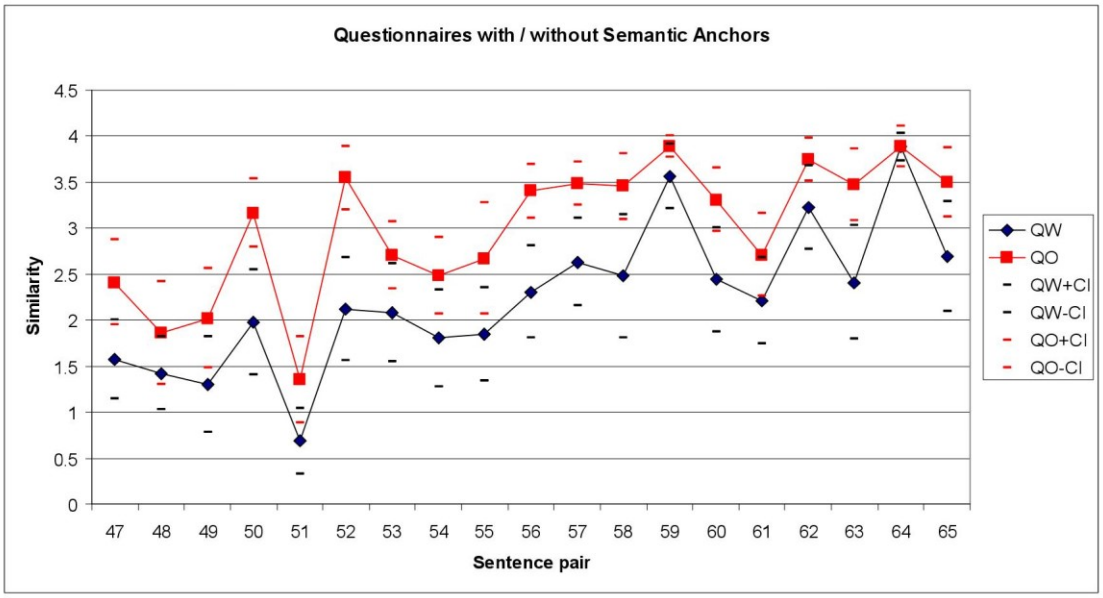
2.10 Analysis of results

The ANOVA analysis indicates whether the Order and Anchor factors have any effect on the measurement of the independent variable, semantic similarity. Having determined that an effect exists, it is then possible to focus on what exactly the effect is. This allows an assessment of the robustness or sensitivity of the experimental technique in revealing the underlying model of STSS – providing evidence of how well the measurement process is getting at the ground truth similarity. Finally it supports the selection of the best combination of levels of the factors for control variables in future studies. Thus the contribution is not only to the experiments in this work, but also the continuing body of work by all researchers in the field.

2.10.1 Effect of varying Order and Anchor

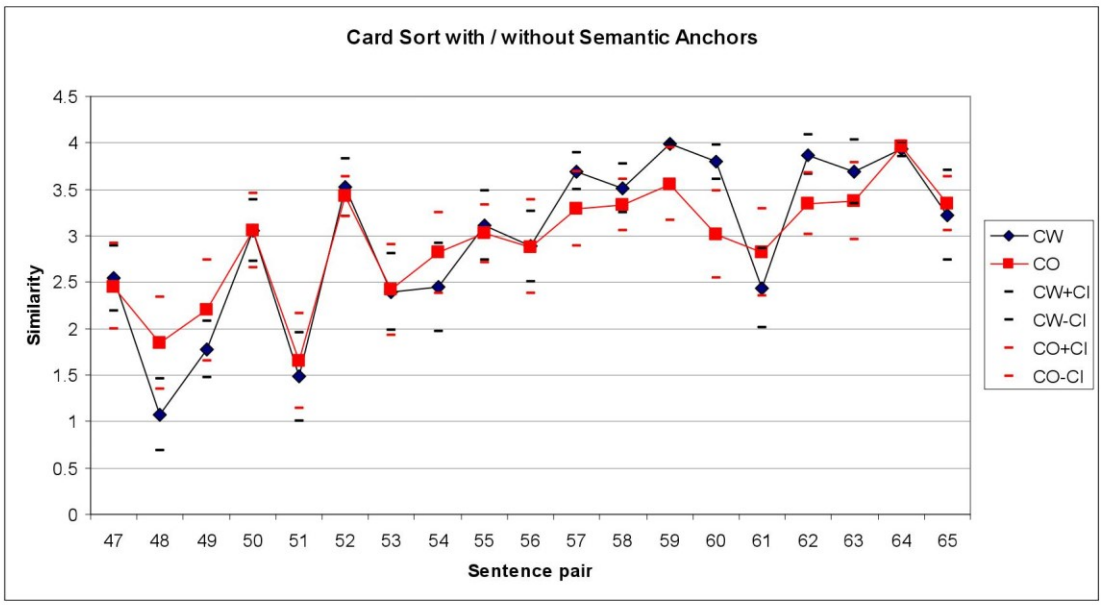
The effect of varying Order and Anchor is best illustrated by line graphs of the medium to high similarity items concentrated in sentence pairs 47-65. Each graph keeps one of the factors constant and plots the trend for both versions of the other factor. The graphs also illustrate the 95% confidence intervals, described in the legend as + or – CI. Figure 2.7 shows the effect of the presence or absence of semantic anchors on the questionnaire form and figure 2.8 shows their effect on the card sort form.

Figure 2.7 The effect of Semantic Anchors on ratings collected using Questionnaires



For the questionnaire form, the two plots are almost identical, except that *without semantic anchors* is displaced upwards to give higher semantic similarity ratings. There is a minor disagreement at sentence pair 49.

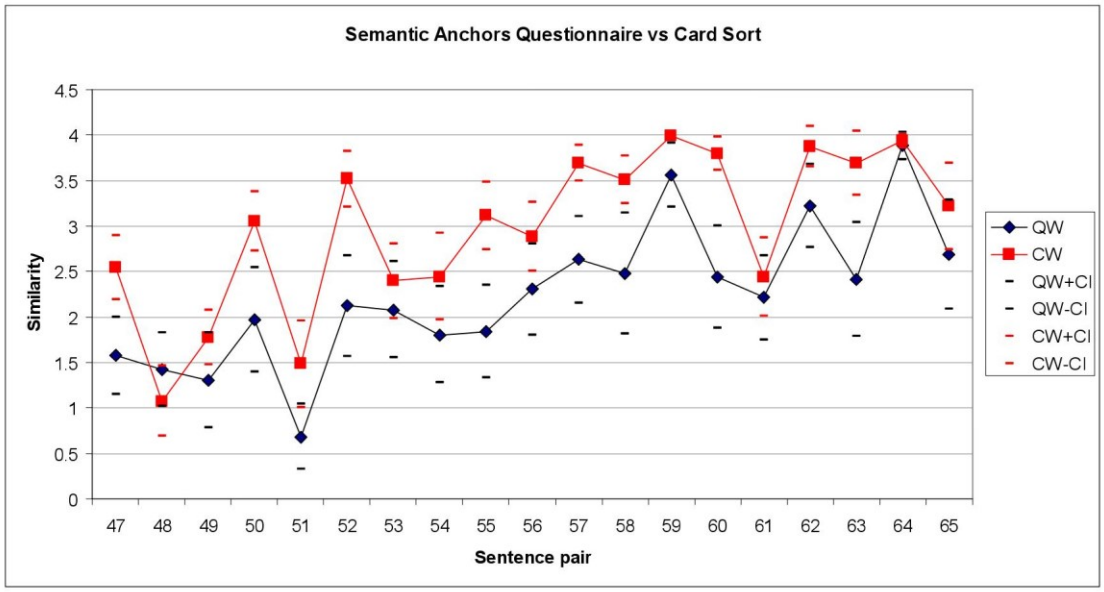
Figure 2.8 The effect of Semantic Anchors on ratings collected using Card Sorting



For card sorting, a number of items have identical ratings. There is more disagreement than the questionnaire form, but there are still regions where the general trends are in agreement. The main disagreements are at sentence pairs 54 and 61.

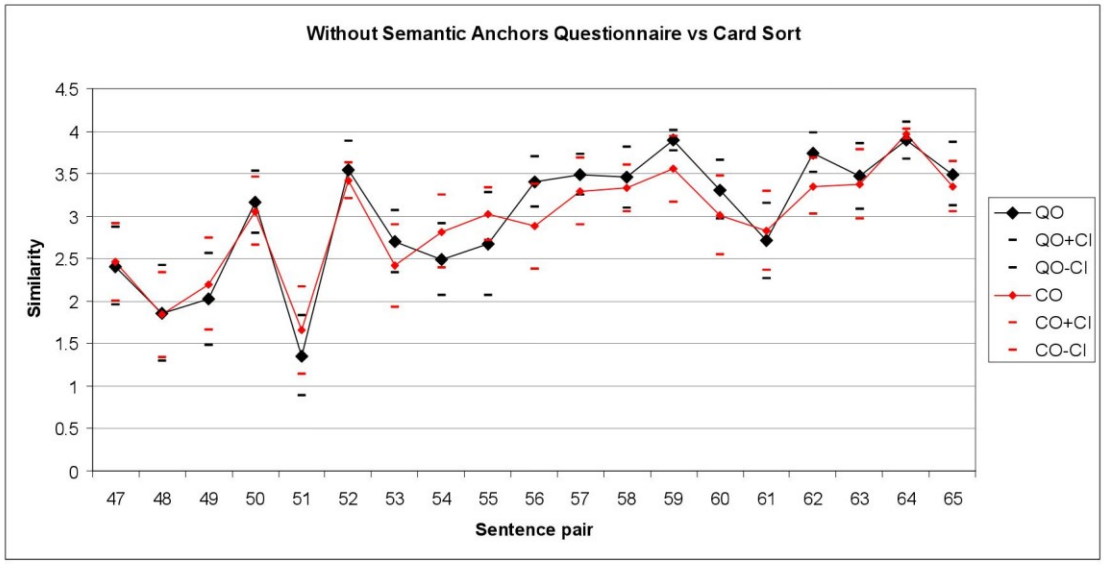
Figure 2.9 shows the effect of using questionnaire or card sorting on the *with semantic anchors* form and figure 2.10 using questionnaire or card sorting on the *without semantic anchors* form.

Figure 2.9 The effect of Questionnaires vs. Card Sorting on ratings collected using Semantic Anchors



For the *with semantic anchors* versions, the overall trends are similar, with the card sorting version displaced upwards to give higher semantic similarity ratings. The noticeable disagreements are at sentence pairs 48 and 57.

Figure 2.10 The effect of Questionnaires vs. Card Sorting on ratings collected without Semantic Anchors



For the *without semantic anchors* versions there are a number of points which are effectively identical and areas with similar trends. The main disagreements are at sentence pairs 54 and 57.

2.10.2 Correlation between experiments

In the following tables, Q means Questionnaires, C means Card Sort, W means With Semantic Anchors and O means withOut Semantic Anchors, e.g. QW means a combination of Questionnaires with Semantic Anchors. As might be expected from the graphs, there is a strong Pearson product-moment correlation between the ratings obtained in each of the experiments. The correlations are shown in table 2.6; the higher the value of r, the better the performance of the measure, with 1.0 indicating perfect agreement.

Experiment	2.1 QW	2.2 QO	2.3 CO	2.4 CW
2.1 QW	X			
2.2 QO	0.959	X		
2.3 CO	0.944	0.981	X	
2.4 CW	0.927	0.981	0.958	X

Table 2.6 Correlations between the ratings obtained from the different experiments

Using 4 groups of 18 participants, all p-values were less than 0.01. In experiment 1 with n = 32, measures of inter-rater agreement (r) were calculated using the leave-one-out method of cross-validation. The average human achieved a correlation of 0.825 and the best human achieved a correlation of 0.921 (with the average ratings for the rest of the group). So the different combinations agree with each other better than the best agreement within the original human group.

2.10.3 Consistency of judgment within experiments

Consistency, which is related to accuracy of judgment, can be investigated by calculating inter-rater agreement using product-moment correlation coefficients (r) as in (Resnik, 1999). The results are shown in table 2.7.

Experiment	Mean r	Best participant r	Worst participant r
2.1 QW	0.855	0.926	0.633
2.2 QO	0.899	0.96	0.75
2.3 CO	0.885	0.919	0.799
2.4 CW	0.938	0.976	0.83

Table 2.7 Inter-rater agreement within experiments

The combination of *Card Sorting with Semantic Anchors* scores best on all three criteria, the correlations of the best and worst humans with the rest of the participants and the average of all of the human participants. The individual correlation coefficients are significant at the 0.01 level.

2.10.4 The effect of the factors on noise

Noise, which is an influence on precision, can be obtained by calculating the standard deviations of the human ratings for each sentence pair within each experiment, then taking their mean, as shown in table 2.8.

Experiment	Mean of SDs
2.1 QW	0.589
2.2 QO	0.723
2.3 CO	0.615
2.4 CW	0.36

Table 2.8 Noise within experiments

The lowest noise value was obtained from the combination of *Card Sorting with Semantic Anchors*. Conducting 2-sample t-tests indicated that there was a strongly significant difference between the noise levels for 2.2 vs. 2.4 and that the difference for 2.3 vs. 2.4 was just on the boundary of significance ($p = 0.05$); the other combinations did not achieve significance.

2.10.5 Effect of order of presentation of sentences within a pair

The data also provides an opportunity to investigate the influence of asymmetry on similarity judgements (Tversky, 1977, Gleitman et al., 1996).

This was investigated using the ratings from the full set of 32 participants originally collected for experiment 1. A sample of 10 sentence pairs (2, 5, 44, 49, 52, 53, 57, 59, 64, 65) spanning the similarity range was selected and the A and B versions were separated out. 2-sample t-tests were conducted to determine whether the means for the two orders of presentation differed significantly. Due to some non-returns of questionnaires there was a small variation in the numbers of A and B versions of questionnaires, therefore the Mann-Whitney test was also conducted for a robust second opinion. The results are shown in table 2.9.

The tests could not be conducted for sentence pairs 2 and 5 because in each case one of the two columns contained only zeros. As virtually all of the data was zeros for these pairs we can assume no effect of ordering.

In all of the other cases, the p-values are substantially greater than the commonly accepted upper limit of 0.05.

Sentence Pair	Human Ratings (0.0-4.0)	2-sample t-test p	Mann-Whitney test p
2	0.02	*	*
5	0.02	*	*
44	0.97	0.191	0.1925
49	1.17	0.144	0.2748
52	1.94	0.626	0.6318
53	1.93	0.694	0.7312
57	2.51	0.898	0.9388
59	3.45	0.126	0.2017
64	3.82	0.929	0.1084
65	2.61	0.366	0.5154

Table 2.9 The effect of order of presentation of sentences within a pair

For both the tests this is considered as providing no evidence for a difference in similarity rating based on order of presentation of sentences within a pair.

2.11 Applications and conclusions

2.11.1 Application to recently developed measures

As an illustration of the value of the dataset, this section draws together self-reported results from 6 studies which have used STSS-65 as a Gold Standard dataset. Most of the approaches follow Li et al in using a corpus distance measure in combination with some other approach. Kennedy and Szpakowitz (Kennedy and Szpakowitz, 2008) use Roget's Thesaurus with a cosine measure, Islam and Inkpen (Islam and Inkpen, 2008) combine a variant of Pointwise Mutual Information (using the British National Corpus) with LCS string matching and Feng et al.(Feng et al., 2008) use Wordnet with the Needleman-Wunsch algorithm to measure indirect relevance. O'Shea et al performed an independent test of LSA using the web portal (Laham, 1998). A comparison of the performance of these algorithms, in terms of the Spearman Product-Moment correlation coefficient (r) with STSS-65, is shown in table 2.10.

These results identify a clear improvement in the performance of STSS measures published in 2008.

Authors	Year	Measure	r
Li et al.	2006	STASIS Wordnet plus word position information in short vectors	0.816
Kennedy & Szpakowitz	2008	Roget's Thesaurus (weighted) plus cosine measure	0.873
Kennedy & Szpakowicz	2008	Wordnet (weighted) plus cosine measure	0.851
Feng et al.	2008	Wordnet and Brown Corpus-based measure incorporating direct and indirect relevance information	0.756
Islam & Inkpen	2008	String matching (LCS) plus SOC-PMI	0.853
O'Shea et al.	2008	LSA	0.838

Table 2.10 The performance of STSS measures

2.11.2 Conclusions

This chapter established that there were no existing STSS datasets and also that there was no quick and easy way to obtain such data. A set of numerical results could be obtained for a set of ST pairs, but it would not be possible to infer that the best process had been used to get to the ground truth or that a valid sample of English STs had been used. Consequently the work appeared to be mired in confounding factors.

The problem was solved by taking a two-stage approach, first (in this chapter) to establish the best procedure, using sentence pairs derived from word-pairs whose semantic similarities were well-known. This established the first benchmark dataset STSS-65. Analysis of the similarities of the sentence pairs and the corresponding word pairs revealed a high correlation between the ST ratings in STSS-65 and the corresponding Rubenstein & Goodenough word pair ratings. Also, explanations could be found for differences in a number of pairs of outliers.

Designing the experimental procedure required a thorough investigation of prior work, particularly in the closely-related field of word semantic similarity. The review of prior work on word similarity revealed a number of strengths and weaknesses, particularly in treatment of experimental controls such as the definition of the measurement scale and sampling the population.

The outcome was that these factors were given careful consideration in designing a process that produced a Gold Standard STSS dataset. The term "Gold Standard" is used somewhat freely but in this case the accolade is deserved on three counts:

- The dataset is the outcome of genuine human ratings rather than some simulation or synthetic procedure
- The dataset was collected using the best available techniques known at the time
- The dataset has been adopted by researchers in the field to validate their algorithms and measure improved performance.

Following the collection of STSS-65 a further set of experiments was conducted to investigate the effect of controlling two additional factors, Order and Anchor. The outcome of these experiments was to validate STSS-65 because the results obtained were very close to the best results obtained in the follow-up experiments.

However, investigation did reveal that one of the combinations of controls, *Card Sorting with Semantic Anchors* was superior to the others; it had significantly lower noise than two of the other combinations and a higher inter-rater agreement than any of the others.

The direction for future work is clear. STSS-65 has proved to be a valuable benchmark dataset, but it has restricted coverage of the language. In particular, all of the STs it contains are statements providing a definition of a word. Also these words are, by and large, concrete nouns.

The requirement is to take the optimal procedure of collecting human ratings, *Card Sorting with Semantic Anchors,* and use it to collect ratings for a more diverse and representative set of ST pairs. These should include instructions, questions etc. which occur in normal dialogue.

The experimental programme in this work is still subject to the constraints of the labour-intensive nature of producing similarity measurements (by participants) and therefore the first challenge of the next chapter will be to determine the best way to represent features of the English language in a restricted set of ST pairs.

This will be followed by experiments to collect and further validate the ratings.

3. Chapter 3 STSS-131 - the new benchmark dataset for STSS

3.1 Introduction

In chapter 2 the need was established for a more demanding dataset to effectively evaluate sentence similarity measures. The new proposed dataset, STSS-131, will overcome the weaknesses identified by being:

- derived from more diverse stimuli than just concrete nouns
- generated by a set of human participants
- more diverse in covering dialogue acts than just assertions

The contributions in this chapter fall into three divisions:

- creation of a new benchmark dataset (STSS-131)
- conducting a discriminablity-based study (semantic distance) on a subset of the items as a validation of the new benchmark dataset and to provide greater insight on the acquisition of human similarity ratings
- evaluating two STSS measures with the new dataset and analysing the degree of challenge compared with the STSS-65

3.2 Creation of new benchmark dataset, STSS-131

A substantial and sound experimental methodology was required to create STSS-131. This broke down into three major phases:

1. Generating a pool of sentences
2. Compilation and selection of a set of 64 sentence pairs from the pool
3. Collecting similarity ratings for pairs of sentences

The design of the process required a number of decisions outlined as follows:

- how to get human participants to generate suitable sentences
- how to represent the language with a limited number of sentence pairs

- how to select a representative sample of sentence pairs
- how to collect the similarity ratings

The method of generating sentences and ratings was adapted from chapter 2.
Instead of using dictionary definitions, human participants were asked to write a sentence they might use in natural dialogue containing a particular stimulus word.
The similarity ratings then used the card sort with semantic anchors method which had been identified in chapter 2 as most suitable.
Representation of the language was achieved by careful choice of stimulus words, described in this chapter. This also contributed to controlling the range of similarities covered by combinations of the generated sentences. Additionally, themes were included for some sentences to promote higher similarity sentence pairings.
Selection of a representative sample of sentence pairs required human judgement. To prevent a biased set of sentences from being used, pooled judgements were used and a short pilot study of similarity ratings was performed before the full study. The following sections will describe the experimental methodology.

3.2.1 Selecting stimulus words

Selecting the stimulus words involved 5 steps:

- choosing the size of the stimulus word set
- decomposing the words of the English language into a hierarchy of classes
- mapping the classes from the decomposition onto a sampling frame
- imposing additional linguistic constraints on certain slots in the frame
- populating the slots in the frame with words with due regard for factors such as word frequency

3.2.1.1 Choosing the size of the stimulus word set

The stimulus word set must be representative of the English language for STSS-131 to test STSS algorithms effectively.
Sentences can be thought of as objects in a semantic space. There are approximately 171,476 words in current use (Coney). Suppose we consider each word position as a point in a multi-dimensional space and each dimension as having 171,476 discrete levels, this is a very large space even after we have eliminated the points in the space which are infeasible sentences (for example where every word in the sentence is the same).

So there is a problem. A large randomly selected sample can represent a population accurately, but this would be prohibitively large. A small randomly-selected sample may be biased by sampling errors and this is known as the representation problem.
Prior word similarity studies have used small samples to represent a large population, without justification of the choices. These range from 6 (Miller and Charles, 1991) to 350 (Finkelstein et al., 2001) word pairings. Most emphasis has been placed on the 48 nouns Rubenstein and Goodenough used to produce their 65 sentence pairs (Rubenstein and Goodenough, 1965).
For this study, the representation problem was solved by adopting a technique, the sampling frame, used by psephologists (Oppenheim, 1992). A sampling frame was constructed, specifying the attributes of each word sampled. Careful control of the attributes meant that even though the sample was small, it was still representative.
Based on the principles (Oppenheim, 1992) and prior work (Rubenstein and Goodenough, 1965) discussed, it was decided to use 64 stimulus words and produce 64 sentence pairs (supporting up to 2016 different combinations if a word can stimulate a sentence for more than one pairing).

3.2.1.2 Decomposing the words of the English language into a hierarchy of classes

Based on general concepts of similarity reviewed in chapter 1 and the consideration of semantic space in section 3.2.1.1, this thesis introduces two properties that can be used in constructing a sampling frame for STSS-131.

- Semantic dispersion - sentence pairs can be considered to be in some way spread throughout the semantic space.

- Semantic proximity – the sentences in a pair can be close (high similarity) or distant (low similarity) in the semantic space.

Consequently, the words making up the English language were decomposed into a tree-structured taxonomy according to particular grammatical, syntactical and semantic properties. Each level is described in the following sections.
Descending through the levels results in increasing numbers of classes, with fewer words, but the words have more semantic commonality. A word in any given class was representative (to some degree) of the class as a whole.

The taxonomy was tree structured, but derived in quite a different way from taxonomies such as WordNet which is based on a small set of operators such as ISA and Part Of (Miller et al., 1990) to perform the decomposition. This prevented biasing the dataset in favour of algorithms that make use of these ontologies.

It was also desirable to represent certain features of the language not related to a particular semantic category (e.g. polysemy, homophony etc.). These features have been incorporated into the frame and are listed as "miscellaneous" at the end of the containing class.

Top level decomposition of the English words

The first decision was to split the whole set of words into the two classes of function and content words (Rubenstein and Goodenough, 1965). Nouns, adjectives, verbs and adverbs are referred to as content words. Other word classes (such as articles, pronouns and prepositions) are referred to as function words. This division is shown in figure 3.1.

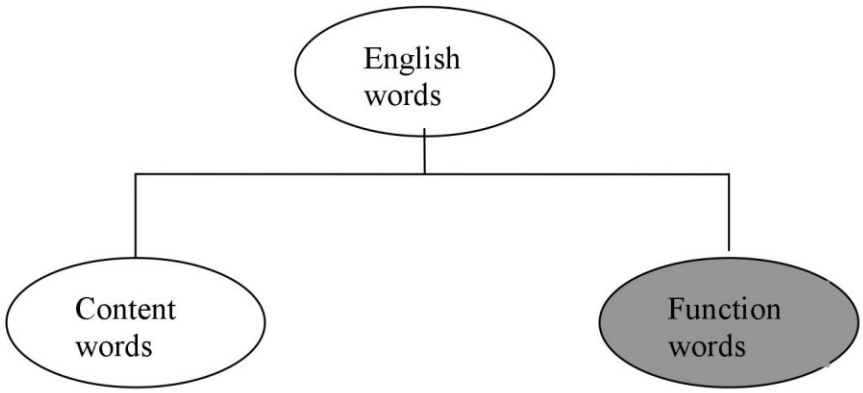


Figure 3.1 Top-level decomposition: Content words vs. Function words

Using content words as stimuli will lead to a natural representation of function words in the generated sentences. For example suppose the stimulus noun *cat* receives the response *The cat sat on the mat*. As a consequence of writing the sentence, the function words *the* (article) and *on* (preposition) have been included as well as the additional content word *mat*. Therefore it was decided to include only content words in the sampling frame.

Decomposition of Content words

The natural consequence of decomposing into content and function words was to further decompose the content words into nouns, verbs adjectives and adverbs. This is shown in figure 3.2

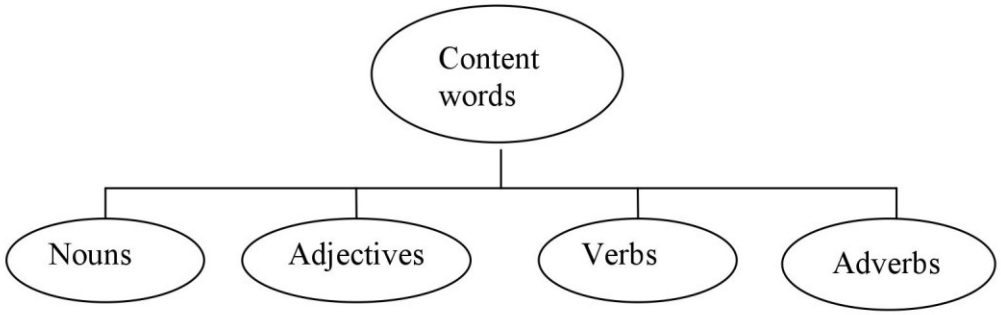


Figure 3.2 Decomposition of the Content words

Word frequency is important in representing the language i.e. words used frequently should have a higher probability of appearing in the sample frame. It was necessary to determine the relative proportions of the occurrences of nouns, adjectives, verbs and adverbs which would occur in the sampling frame. Corpus statistics provide the best source of frequency information. There are a number of corpora in existence - the Brown Corpus, Penn Treebank, Collins Cobuild and the British National Corpus (BNC) are all familiar in the literature. The Penn Treebank was eliminated because of its specialised nature (materials drawn from the Wall St. Journal). The Collins Cobuild was eliminated because there were significant charges for use. This left the Brown Corpus and the BNC, both derived from a range of printed materials, both freely available to researchers and both widely used and understood by researchers. Brown has been established longer but the BNC is larger (at least 100 million words (Webmaster)) and more up-to-date.
The problem of representing the frequencies was solved by distributing the words in the frame to word classes based on their frequency of occurrence in the two corpora. This was achieved by summing the total number of occurrences of all of the words (lemmata in the case of the BNC) in each class and finding the average for the two corpora. These were then used to weight the proportions in the frame. Figure 3.3 shows the results of distributing the 64 words in these proportions.

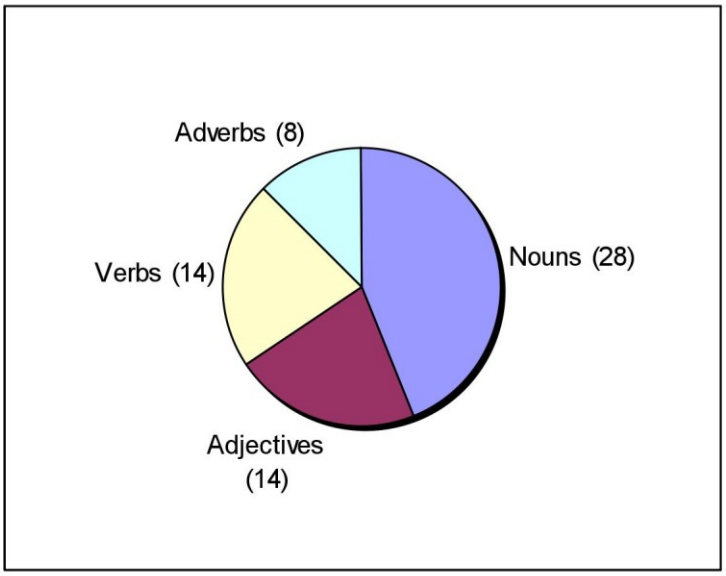


Figure 3.3 The distribution of 64 word slots between the Content word classes

Representation of frequency within the Content word classes

Within each class the words will also follow a frequency distribution. For valid representation of the English language, words selected for the class should be in proportion to their frequency of occurrence.

There are a number of approaches to selecting words on frequency, for example the biased roulette wheel as used in Genetic Algorithm cost functions (Goldberg, 1989) was considered.

However, there are known problems with word frequency lists below the frequencies required to make it into the first five or six thousand words (Francis and Kucera, 1982, Juola, 1996), so a simpler and more robust approach was chosen.

This was to split the English language into two divisions – high frequency and low frequency words. A number of researchers have compiled lists of high frequency words over the years (Ogden, 1930, West, 1953, Coxhead, 2000).

The motivation has been to identify a "most useful" set of English words to use in language teaching. Most attention focuses on the idea of a core vocabulary, which contains the most useful and frequently-used 2,000 words of English. It has also been reported that over 80% of the words in Undergraduate textbooks are members of a basic high frequency word set (Valcourt and Wells, 1999).

The problem of frequency representation within the classes was solved by deciding to split the words in the classes into 80% coming from a high frequency word list and 20% coming from a low frequency word list. The high frequency word list was produced by combining

the most frequent 2,000 words from each of the Brown and BNC corpora. The low frequency list would come from the remainder of the BNC after the words on the high frequency list were removed.

Consequently, it was decided to set a target of 22 high frequency and 6 low frequency nouns, 12 high frequency and 2 low frequency adjectives, 11 high frequency and 3 low frequency verbs and 6 high frequency and 2 low frequency adverbs. The decomposition by frequency is shown in table 3.1.

The final column of table 3.1 shows the breakdown when each class is split in the proportions 80:20.

	Brown Corpus	BNC Lemmatised List	Words in Class	Frequency Breakdown
Nouns	31	25	28	22 High
				6 Low
Adjectives	20	7	14	12 High
				2 Low
Verbs	8	22	14	11 High
				3 Low
Adverbs	6	10	8	6 High
				2 Low
Totals	65*	64	64	51 High
				13 Low

Table 3.1 Frequency breakdown for English Content Words

* The Brown corpus figures add up to 65 instead of 64 because of variations introduced by rounding.

Each of the classes was then subject to further decomposition separately, starting with the nouns.

Decomposition of the nouns - methodology

A number of possibilities were considered for decomposing the nouns further. These included ontological approaches such as WordNet (Miller, 1995, Miller et al., 1990) and Thesauri (Davidson, 2004), grammatical approaches and Category Specific Deficits (CSDs) introduced in chapter 1.

The decision was made not to proceed with Ontological approaches because some existing STSS measures already use Wordnet or Roget's Thesaurus in calculating semantic similarity (Li et al., 2006, Kennedy and Szpakowitz, 2008, Feng et al., 2008). It was important to avoid biasing the dataset in favour of these algorithms.
Consequently, it was decided to proceed with a combination of higher-level decomposition using conventional grammar followed by lower level decomposition using CSDs.

The Concrete / Abstract division

English grammar offers provides a way of decomposing the nouns into further intermediate categories. The first of these is between Concrete and Abstract.
A concrete noun (e.g. brick, screwdriver) labels an object which can be interacted with directly through the senses, whereas an abstract noun refers to some sort of intangible quality (e.g. truth, beauty). Nouns are sometimes described as having a "degree" of concreteness dependent on the extent to which they can be experienced directly by the senses (Bushell and Martin, 1997).
A surprising number of abstract nouns appear in the most frequent 2,000 words. Searle (Searle, 1999) suggests that the ability to form abstract concepts (such as money and government) and manipulate them is fundamental to our success as a species. For example:

> "*Money* earned as an *employee* of the *State of California* in his *bank account* which he uses to pay *state and federal taxes* as well as *bills* owing to the *gas and electric companies* and to the *contractor* of his *credit cards.*"

Abstract nouns also pervade science. Things that we can measure quantitatively nevertheless use scales with abstract properties, e.g. *temperature*.
The concrete / abstract division is also supported by evidence from CSDs (Damasio et al., 2004, Breedin et al., 1994, Newton and Barry, 1997). The division is illustrated in figure 3.4.
Consequently, it was decided to reserve 8 slots in the sampling frame for abstract nouns. Of the remainder, 19 were allocated to concrete nouns and one reserved for special language properties, discussed later.

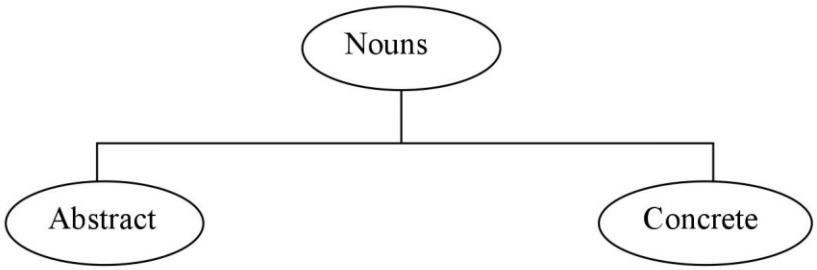


Figure 3.4 Decomposition of the Nouns

Decomposition of the Abstract Nouns

The final stage for abstract nouns was to decompose them into individual bottom-level classes, each of which contained a word in the sampling frame. Some useful classes of abstract nouns are available in definitions from dictionaries and language teaching resources, for example:

- Qualities or ideas(Sinclair, 2001)
- Feeling, ideas and qualities (Lisa, 2009)
- States, events, concepts, feelings and qualities (Flynn, 2002a)
- Emotions/Feelings, States/Attributes, Ideas/Concepts/Ideals and Movements/Events (Yourdictionary, 2009)

Given the importance of science in society, it was also decided to specify a specific "idea from science" as well as a general idea. Also the attribute "affect" is generally important in the psychological study of the subjective aspect of feeling or emotions. Therefore the emotion slot was spilt into two, one for positive, the other for negative, affect.
Finally, it was felt that where there were sufficient slots for a particular class, some completely random allocation would be beneficial. In the case of Abstract nouns one slot was available for this purpose.
Consequently, it was decided to include the following abstract noun slots in the frame:

- Quality
- Idea (from Science)
- Institutional Fact
- Idea (randomly selected)
- State

- Emotion positive
- Emotion negative
- Randomly selected

Decomposition of the concrete nouns

Concrete nouns can be Proper, Common or Collective (Thomson and Martinet, 1969). A Proper noun is a name (for example Tom, Paris) and a Collective noun refers to a mass of items as a whole (crowd, pile). A Common noun is a concrete noun which is neither proper nor collective (e.g. desk, tree).

The semantic properties of names are likely to be experiential arising from the judge's life history.

There is little information available about the relative frequencies of collective vs common nouns. However, examination of web lists suggests somewhere between 50 -100 collective nouns in genuine use throughout the language.

Consequently it was decided to omit proper nouns from the sampling frame and split the 19 dedicated concrete noun slots into 17 common and 2 collective slots (one for living, the other for nonliving), as shown in figure 3.5.

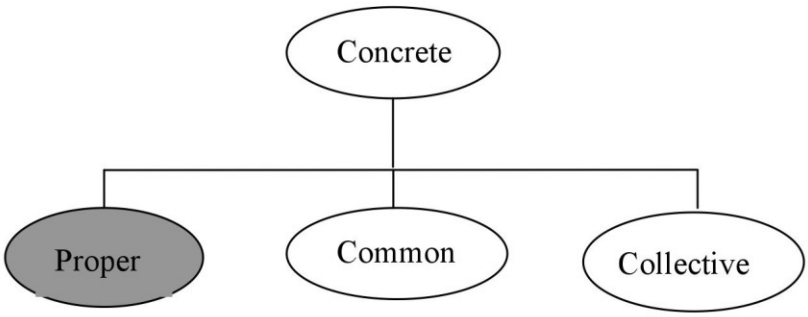


Figure 3.5 Decomposition of the Concrete Nouns

CSD-based decomposition of the Common and Collective nouns

CSDs offer the opportunity to decompose to the final semantic categories. Concrete nouns undergo a two stage process, whilst collective nouns undergo a single stage decomposition. The processes are described in the following sections.

Intermediate decomposition of the nouns using CSDs

After careful consideration, the decision was taken to use CSD information for the later stages of decomposition. There is substantial evidence from CSD studies to support a division of nouns into two classes, which is applicable to both common and collective nouns, in different ways, after they have been split.

There are three different terminologies for division into two classes: Living / Nonliving (Pouratian et al., 2003), Biological / Nonbiological (Vinson et al., 2003) and Animate / Inanimate (Caramazza and Shelton, 1998). This poses the question: are these simply terminological differences describing exactly the same division into two classes or are divisions subtly different from each other?

The living/nonliving dichotomy was used in Vinson (Vinson et al., 2003). Examples of living things were animals (tiger, lion), faces, fruits and vegetables. Examples of nonliving things were furniture, kitchen utensils, clothes and vehicles. This study found that body parts had more in common with nonliving things than living things and treated them as such. Other studies on the living/nonliving dichotomy include (Moss and Tyler, 2000, Devlin et al., 2002, Cross et al., 2008, Lu et al., 2002). The living/nonliving dichotomy also appeared in a study of semantic priming in an ERP study using healthy participants (Sartori et al., 2006).

A study which used 3 different datasets to investigate the statistical methodologies involved in CSD research (Laws et al., 2004) made a distinction between living things (words from fruits & vegetables and animals categories, e.g.mushrooms, vulture) and objects (e.g. darts, passport).

The "inanimate" term appears in a review of prior art in (Caramazza and Shelton, 1998) where some work is described as using living/nonliving catgeories and other work as using living/inanimate categories. Animals are commonly considered as living things, but the fruits and vegetables are treated differently in the reviewed studies. Thus fruits and vegetables are considered living in one study and inanimate in another. This distinction is in fact consistent with dictionary definitions (Little et al., 1983, Sinclair, 2001) of animate which generally focus on an entity having animal life and also with connotations of motion. Throughout, the nonliving or inanimate items are described as "artefacts" (Caramazza and Shelton, 1998) without specific examples. In the experimental section of the study, the two major categories used are animals and nonanimals, with nonanimals including body parts, clothing, fruit, furniture, kitchenware, musical instrument, tool, vegetable, vehicle and other.

In an exchange of letters in Trends in Cognitive Sciences (Mahon and Caramazza, 2003), Mahon and Caramazza describe the domains living animate, living inanimate and nonliving as specific categories that can each be disproportionately impaired. A contemporaneous paper including some of the same authors uses the terms animate objects, inanimate biological objects and artefacts (Capitani et al., 2003); this study also describes biological objects as dividing into the two categories animal and fruit & vegetables.

Biological and living are sometimes treated as the same thing. This is the case in (Kay and Hanley, 1999), which gives example subcategories of animals, birds, fruits, vegetables. It also specifies the alternative category as inanimate and provides example subcategories of furniture, kitchen utensils and clothes. The terms artefact and nonliving are used in other parts of the paper for the inanimate items.

Both biological and living are used as category descriptors in a review of impairments from Alzheimer's disease reviewed by Whatmough (Whatmough and Cherkow, 2002).

The problems of allocating the categories of food and body parts to the higher level domains of living and nonliving is noted by Gainotti and Silveri (Gainotti and Silveri, 1996), who observed food is usually impaired along with living things, whereas body parts are usually spared, despite the fact that much food is manufactured (processed) and unrecognisable in terms of the features of its living origins, whereas body parts are usually components of a living being. Finally, Chao (Chao et al., 1999) used the terms biological and nonbiological consistently in an fMRI study of categories in healthy volunteers. The biological categories were animals and faces, the nonbiological categories were tools and houses.

Consideration of the above and other prior work led to the decision that for the work in this thesis classes would be formed using both living/nonliving and biological/nonbiological distinctions.

Therefore, the four possible classes that emerged were *living/biological*, *nonliving biological*, *living/nonbiological* and *nonliving/nonbiological*. These would provide a good medium-grained partitioning for concrete nouns and would contain examples of the domains of animate/inanimate within their subcategories. Body parts would be included under *living/biological*, as would fruits and vegetables, but processed materials such as Spam (canned pork and ham) and cotton would be included under *nonliving/biological*.

It is debatable whether the *living/nonbiological* category has any genuine members (e.g. robot, android and virus could also appear in the nonliving biological category (Rice, 2006)). Consequently it was decided to omit *living/nonbiological and* split the common

nouns into the classes *living/biological*, *nonliving/biological* and *nonliving/nonbiological.* This decomposition is shown in figure 3.6.

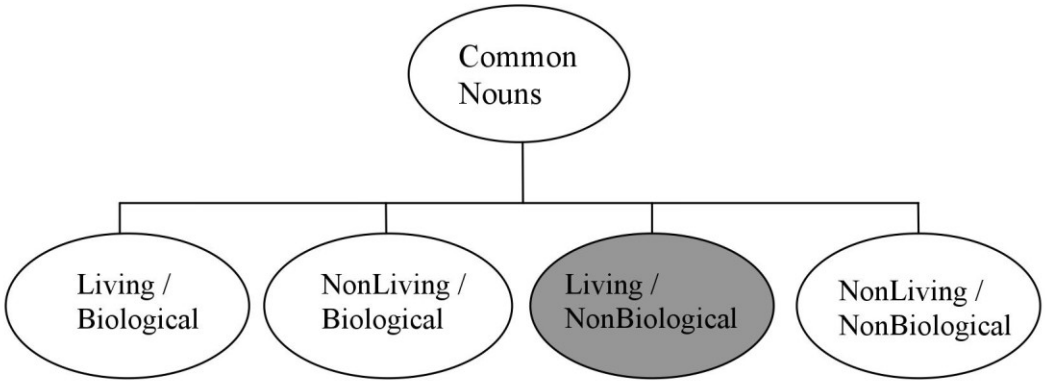


Figure 3.6 Decomposition of the Common Nouns

At this stage the numbers of noun slots were dominated by the availability of established classes from CSD studies. The 17 common noun slots were divided into 8 *living/biological*, 8 *nonliving/nonbiological* and 1 *nonliving/biological* (which has a relatively minor role).

Final selection of classes of common nouns

The final stage of decomposition of the common nouns used established classes from CSD studies. Table 3.2 shows the classes chosen. The criterion for selecting the classes was one of credibility - following peer review or repeating of experiments. **Bold type** indicates multiple sources confirming a particular neuroanatomical site linked to the class and the capital R indicates evidence from a large scale review of other work.

Column 4 identifies some of the neuroanatomical structures associated with particular categories. Some researchers use a naming system in which major temporal structures are identified with their initial capital letter. For example O-T refers to Occipital-Temporal. These are illustrated in figure 3.7 (Serendip, 2003). Capital I refers to Inferior.

Gross Category	Specific category	Sources	Identified Neuroanatomical structures
Living / Biological	Body Parts	(Rogers and PLaut, 2002) R (Vigliocco et al., 2002a) (Capitani et al., 2003) R (Vigliocco et al., 2002b) (Gainotti and Silveri, 1996) (Forde et al., 1997)	Lateral fusiform gyrus (Vigliocco et al., 2002a)R Bilateral inferior T, O-T, fusiform, lingual
Living / Biological	Fruits and Vegetables	(Vigliocco et al., 2002a) (Capitani et al., 2003) R (Damasio et al., 2004) (Santos and Caramazza, 2002) R (Forde et al., 1997)	**(Damasio et al., 2004)** Bilateral inferior T (Capitani et al., 2003)
Living / Biological	Animals	(Rogers and PLaut, 2002) R (Caramazza and Shelton, 1998) (Vigliocco et al., 2002a) (Damasio et al., 2004) (Forde et al., 1997)	Posterior ventral temporal cortex, prefrontal motor areas (Rogers and PLaut, 2002) **Occipital-Temporal areas (Vigliocco et al., 2002a) R** Medial areas(Vigliocco et al., 2002a) R **(Damasio et al., 2004)**
Living / Biological	Birds	(Capitani et al., 2003)R (Forde et al., 1997) (Forde and Humphreys, 2002)	Left F-P (intermediate impairment for birds) (Capitani et al., 2003)
Living / Biological	Insects	(Capitani et al., 2003)R (Forde et al., 1997) (McRae and Cree, 2002)	Bilateral T-P atrophy Inferior T-O (Capitani et al., 2003)
Living / Biological	Plants and Flowers	(Rogers and PLaut, 2002) R (Caramazza and Shelton, 1998) (Capitani et al., 2003)R	Left F-T and insula, right T insula (with other biologicals) (Capitani et al., 2003)
Nonliving / Biological	Foodstuffs	(Rogers and PLaut, 2002) R	Bilateral T **(Capitani et al., 2003)**

		(Capitani et al., 2003) R (both naturally occurring and manufactured) (Gainotti and Silveri, 1996) (Santos and Caramazza, 2002)	
Nonliving / Nonbiological	Clothing	(Rogers and PLaut, 2002) R (Cloth types) (Vigliocco et al., 2002a) (Capitani et al., 2003)R (Vigliocco et al., 2002b) (Forde et al., 1997)	Left t and basal ganglia (with other non-animal items) Left T-P (fabrics) (Capitani et al., 2003)
Nonliving / Nonbiological	Tools and Manipulables	(Vigliocco et al., 2002a) (Capitani et al., 2003) R (Damasio et al., 2004) (Forde et al., 1997)	**Left Posterior Middle Temporal Gyrus (Devlin et al., 2002) (Damasio et al., 2004)** Middle temporal gyrus (Rogers and PLaut, 2002) Lateral Temporo-Parietal-Occipital (Vigliocco et al., 2002a) R Lateral Temporal Cortex (tool motion) (Vigliocco et al., 2002a) R
Nonliving / Nonbiological	Vehicles	(Vigliocco et al., 2002a) (Capitani et al., 2003) R (Vigliocco et al., 2002b) (Santos and Caramazza, 2002) R (Forde et al., 1997)	Left T and Basal Ganglia (with other non-animal categories) Mild cerebral atrophy (Capitani et al., 2003)
Nonliving / Nonbiological	Furniture	(Vigliocco et al., 2002a) R (Capitani et al., 2003)R (Santos and Caramazza, 2002) R (Forde et al., 1997)	(Chairs) Inferior Temporal Gyrus (Vigliocco et al., 2002a) Left anterior T Left T-P (Capitani et al., 2003)
Nonliving / Nonbiological	Musical Instruments	(Rogers and PLaut, 2002) R (Capitani et al., 2003) R (Damasio et al., 2004)	Left Temporal pole and Anterior Ventral I-T (Damasio et al., 2004)

		(Gainotti and Silveri, 1996) (Miller and Charles, 1991)	
Nonliving / Nonbiological	Miscellaneous Artefacts	(Rogers and PLaut, 2002) R (Capitani et al., 2003) R	**(artefacts) Left Middle Temporal Gyrus Dorsolateral Frontal Cortex (Vigliocco et al., 2002a) R**
Nonliving / Nonbiological	Gemstones	(Rogers and PLaut, 2002) R (Warrington and Shallice, 1984) (McRae and Cree, 2002)	

Table 3.2 Semantic categories for nouns selected from the field of neuropsychology

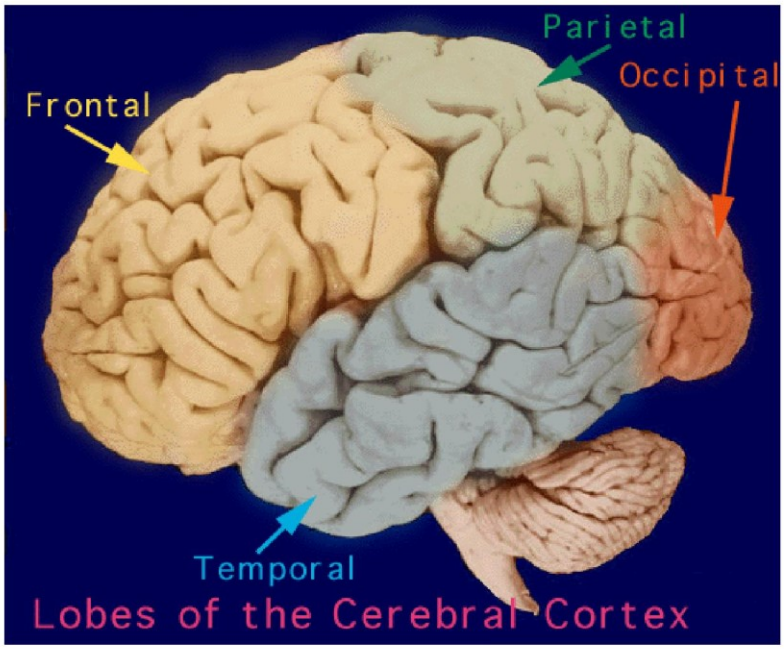


Figure 3.7 Lobes of the Cerebral Cortex

Decomposition of the collective nouns

A pure collective noun is a term for a collection of entities such as “flock” of sheep. Collective nouns are sometimes confused with mass nouns which denote a quantity of some substance such as “water”. Semenza defines collective nouns as being "like flock or furniture that denote groups of elements that cannot be enumerated in each single component but may or may not share the feature of countability (Semenza et al., 1997). i.e. it is possible for a farmer to own 3 flocks (of sheep) but a householder is never described as owning 3 furnitures. As stated earlier, collective nouns may be decomposed into the same intermediate categories as common nouns, but for practical purposes it was decided

to restrict the classes to Living/Biological and Nonliving/Nonbiological. Consideration of the nonliving biological class suggested that their collective terms (such as "bunch") were ambiguous. This is shown in figure 3.8.

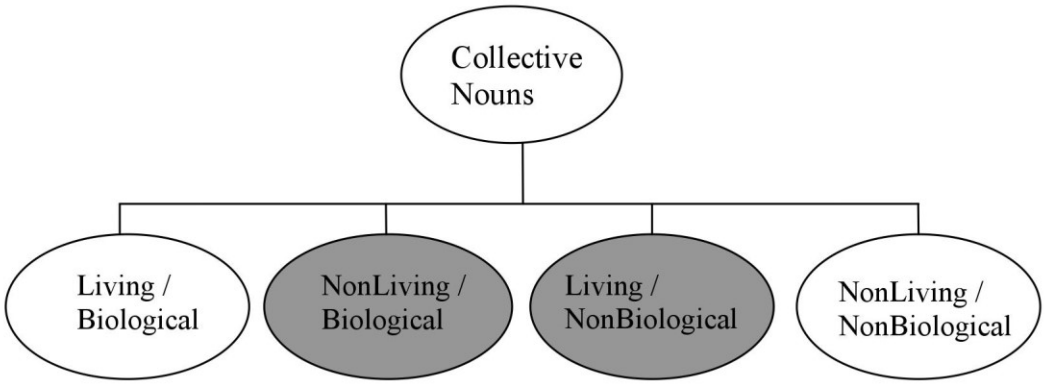


Figure 3.8 Decomposition of the Collective Nouns

Miscellaneous noun slots

It was decided to reserve two slots in the sampling frame, one for a Living / Biological and the other for a Nonliving / Nonbiological collective noun.

Two slots were reserved to promote diversity of coverage (valid semantic classes may exist which are not recorded in CSD literature). These were:

Any word that fitted the Living / Biological class, regardless of its sub-class

Any word that fitted the Nonliving / Nonbiological class and was not in the man-made class (ass this already had good representation).

Certain properties of words such as polysemy and homophony have long been of interest to etymologists and may have a role to play in STSS. Although it was possible that some of these properties would be represented naturally as part of the selection process, it was decided to build-in at least some representation at this stage. The properties considered were:

- Polysemy: a polysemous word has a number of distinct meanings, but the meanings are related
- Homonymy: a homonym has the same pronunciation and possibly the same spelling as another word with a different meaning
- Homophony: a homophone has the same pronunciation as, but a different spelling and meaning from another word with a different meaning

- Heteronymy: a heteronym has a different pronunciation, but the same spelling as another word with a different meaning
- Homography: a homograph sometimes has the same pronunciation and always has the same spelling, as a word with a different meaning (and etymological origin)
- Antonymy: an antonym has the opposite meaning to another word.

In English virtually all high-frequency words are polysemous to some extent, so some will be included in the sampling frame automatically. Therefore polysemy was eliminated from needing representation.

Clearly there is scope for overlap between the definitions as the pair *knight* and *night* qualifies as both homonym and homophone. Also, the pair *row* (a boat) and *row* (argue) qualifies as both heteronym and homograph.

Given the space constraints it was decided not to try and represent the fine distinctions between these properties.

Consequently, it was decided to reserve one slot in the nouns for a homonym or homograph of one of the other nouns (which of these would be determined by a technique for populating the frame described later).

As discussed earlier, in the cases of both the homonym/homograph and the antonym if the requirement had already been satisfied by the sampling process, then the slot could be released for random allocation of a noun or adjective, respectively.

A relationship described as a homonymous noun-verb pair (Viggliocco et al., 2005) was accommodated by adding a constraint that one of the nouns must be part of a homonymous pair with one of the verbs (e.g. *fight* can be both a verb and a noun).

Final structure of the noun slots

The final structure of the noun section of the sampling frame is presented in table 3.3.

Slot	Grammatical Class	Sub-class
1	Abstract	Quality
2	Abstract	Idea from Science
3	Abstract	Institutional Fact
4	Abstract	Idea
5	Abstract	State
6	Abstract	Emotion positive
7	Abstract	Emotion negative
8	Abstract	Randomly selected
9	Concrete:Common	Living:Biological:Body Parts (CSD)
10	Concrete:Common	Living:Biological:Fuits and vegetables (CSD)
11	Concrete:Common	Living:Biological: Animals (CSD)

12	Concrete:Common	Living:Biological: Birds (CSD)
13	Concrete:Common	Living:Biological: Insects (CSD)
14	Concrete:Common	Living:Biological: Plants and Flowers (CSD)
15	Concrete:Common	Nonliving:Biological: Foodstuffs (CSD)
16	Concrete:Common	Nonliving:Nonbiological:Clothing (CSD)
17	Concrete:Common	Nonliving:Nonbiological:Tools and Manipulables (CSD)
18	Concrete:Common	Nonliving:Nonbiological:Vehicles (CSD)
19	Concrete:Common	Nonliving:Nonbiological:Furniture (CSD)
20	Concrete:Common	Nonliving:Nonbiological:Musical Instruments (CSD)
21	Concrete:Common	Nonliving:Nonbiological:Miscellaneous Artefacts (CSD)
22	Concrete:Common	Nonliving:Nonbiological:Gemstones (CSD)
23	Concrete:Common	Nonliving:Nonbiological:Other – Any non man-made (CSD)
24	Concrete:Collective	Living:Biological (CSD)
25	Concrete:Collective	Nonliving:Nonbiological (CSD)
26	Concrete:Common	Living:Biological:Any sub-category (CSD)
27	Concrete:Common	Nonliving:Nonbiological: Other - any sub-category (CSD)
28	*Homonyn/Homophone	Of any previously selected word or Randomly selected

Table 3.3 Noun Classification sub-Frame

The process of selecting words to fill the slots in the frame will be described later in this chapter.

Decomposition of the adjectives

As shown in table 3.1, 14 slots in the sampling frame were reserved for adjectives. This section explains how they were allocated.

Adjectives have not been used in CSD studies and no other alternatives to grammatical methods of decomposing them were found. Also because fewer adjectives were allocated as stimulus words than nouns, a smaller number of coarser categories was required.

At the top level, a formal grammar text may divide adjectives into 6 categories (Thomson and Martinet, 1969):

- Quality (e.g. heavy)
- Demonstrative (e.g. this)
- Distributive (e.g. each)
- Quantitative (e.g. some)
- Interrogative (e.g. which)
- Possessive (e.g. my)

Intuitively most people associate adjectives with adjectives of quality. Indeed all of the examples of non-qualitative adjectives (*this, each, some, which* and *my*) are also members of the class of function words. Therefore the top-level breakdown of the adjectives is as shown in figure 3.9.

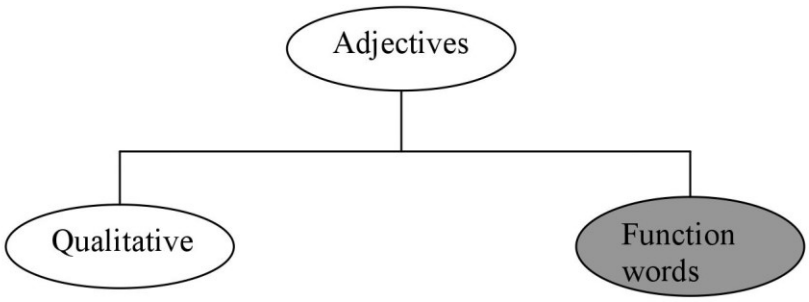


Figure 3.9 Breakdown of the Adjectives

As discussed earlier in this section, those adjective classes which consist solely of function words need no further consideration. One of the best-established methods of decomposing adjectives, Dixon's, is derived from English(Dixon, 1991). Dixon's principles have also been proven across a range of languages (Dixon and Aikhenvald, 2004). Although there has been substantial work in decomposing German adjectives (Naumann, 2000)}(Lee, 1994, Hundsnurscher and Splett, 1982. , Rachidi, 1989), the classes were not directly applicable.

Consequently it was decided to adopt Dixon's typology for decomposing the qualitative adjectives.

Dixon's typology contains 10 classes, from which an eleventh was derived in order to represent affect (Rossell et al., 1988):

- Dimension
- Physical property
- Colour
- Age
- Value (positive affect)
- Value (negative affect)
- Speed
- Human Propensity
- Similarity

- Difficulty
- Qualification

It was also decided to place a restriction on the human propensity slot. Human propensity is a disposition to behave in a certain way. Because of the importance placed on personality and emotion in psychology (McCrae and Costa, 1987, Saucier and Goldberg, 1996, Benet-Martinez and Waller, 2002, Hassebrauck and Buhl, 1996), it was decided to dedicate this slot to an "evaluative personality descriptor".

Miscellaneous adjective slots and constraints

The remaining 3 slots from the 14 were allocated to specific areas of interest that arose during the literature search:

- Smells
- Degree
- Antonymy

Smells have a unique relationship with memory, known as the "Proust Phenomenon" (Zucco, 2003, Herz et al., 2004), (Lawless and Cain, 1975) and brain physiology (Puppala et al., 2009) (Bédard and Parent, 2004). Consequently a slot was reserved for an adjective representing a smell.

Adjectives have a comparison property, i.e. they can express a degree of the attribute. Consequently a slot was reserved for an adjective in the comparative form.

Some classes of adjectives support the properties of antonymy (oppositeness), for example the physical properties *hot* and *cold* and the dimensional properties *long* and *short* may be considered opposite. Antonymy is of interest in semantic similarity because co-occurrence measures of word pair similarity tend to give high scores to antonymous pairs (Rubenstein and Goodenough, 1965, French, 2001). Consequently a slot was reserved for an antonym of one of the other adjectives in the frame.

It is also possible for adjectives to be part of a homonymous verb-adjective pair (e.g. *smooth*) so a constraint was added that these should also be represented.

The final format of the section of the adjective sampling frame is presented in table 3.4.

Adjective classification sub-frame

Slot	Class	
1	Dimension	
2	Physical property	
3	Colour	
4	Age	
5	Value	Positive Affect
6	Value	Negative Affect
7	Speed	
8	Human Propensity	Evaluative personality descriptor
9	Similarity	
10	Difficulty	
11	Qualification	
12	Smell	
13	Any class except 5/6/	Comparative (-er form)
14	Any class except 5/6/	Antonym

Table 3.4 Adjective Classification sub-Frame

Decomposition of the verbs

As shown in table 3.1, 14 slots in the sampling frame were reserved for verbs. To meet the requirement of breaking them down into intermediate and finer-grained classes, the following approaches were considered:

- CSDs
- Traditional grammar
- Lexical construction approaches
    - Lexical Conceptual Structure
    - Role and Reference Grammar
    - Case Grammar
    - Levin Classes

Category Specific Deficits

For verbs, CSDs do not provide fine-grained semantic classes; although they provide independent objective evidence for some of the higher-level distinctions made by the grammatical approaches (Silveri et al., 2003 , Miller, 1972, Bushell and Martin, 1997, Lu et al., 2002, Vigliocco et al., 2004).

Consequently it was decided not to make use of CSDs in this section.

Traditional Grammar

Traditional grammar provides 4 classes that are useful in a high-level decomposition: Primary Auxiliaries, Modal Auxiliaries, Catenatives and Full Verbs (Palmer, 1988).

- **Primary Auxiliaries**

  The primary auxiliaries are a small set containing the verbs *be, have* and *do*. They are fundamental to the construction of tenses and questions.

- **Modal Auxiliaries**

  Modal auxiliaries contain verbs such as *would, may, ought* etc. They are frequently used in constructions to express possibility, desire or obligation.

- **Catenatives**

  Catenatives are verbs (other than the auxiliaries), which may be chained together to form extensive verb phrases, for example: *I don't want to have to be forced to begin to try to make more money* (Palmer, 1988).

- **Full Verbs**

  Full verbs are all the remaining verbs after the auxiliary and catenative classes have been removed.

All of the verbs in the primary and modal auxiliary classes are function words, therefore they are not required in the sampling frame.

There are many catenatives in ordinary use and because they share the properties of full verbs it was decided not to treat them as a separate class.

Consequently it was decided to select from a list of verbs with the auxiliaries removed and assume that catenatives would be included naturally. This is shown in figure 3.10. From this point on the term "full verbs" will be taken to include catenatives.

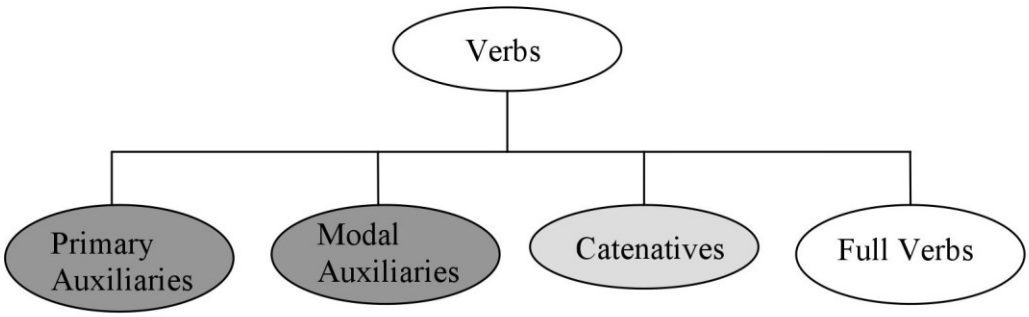


Figure 3.10 Top-level decomposition of the Verbs

Further properties of traditional grammar include properties such as the transitive / intransitive distinction. For example "cut" is used transitively in "The butcher cuts the meat." And intransitively in "The meat cuts easily." Any single distinction does not help to separate verbs into semantic classes. Therefore an alternative approach was required at this stage and this was found in a set of NLP techniques known as "Lexical Construction."

Lexical Construction

Lexical construction emerges from more sophisticated grammatical approaches developed for NLP. This umbrella term covers the motion-based approach of Lexical Conceptual Structure (LCS), grammar-based approaches of Role and Reference Grammar (RRG) and Case Grammar (CG), and Levin's Alternation system (Levin, 1993).

Lexical Conceptual Structure

The LCS classification system was formulated by Jackendoff (Jackendoff, 1983, Jackendoff, 1990), building upon earlier template-based approaches and observational studies. LCS uses patterns e.g.[$_{event}$ GO$_{+loc}$ ([$_{thing}$], [$_{path}$]) and verbs are classified based on their subsumption by generic LCS patterns (Drakos, 1998).

Problems with LCS are the complexity of applying the template-based approach and the fact that it produces a flat structure with no intermediate decompositional levels (Drakos, 1998). Also, according to Traum (Traum and Habash, 2000) the semantic structure of the verb classes duplicates Levin's typology.

Consequently it was decided not to use LCS for verb class decomposition.

Role and Reference Grammar

The grammar-based approaches usually specify a split into State and Non-State verbs at the top level. RRG differentiates between a state class and 3 non-state categories (achievements, accomplishments and activities), however the non-state categories are derived from the state class(Jackendoff, 1983). Thus learn is derived from the state *know* as BECOME know' and *teach* is also derived from *know* as CAUSE [BECOME know']. Although the state verbs offer 6 interesting categories (locational, perception etc.) (Van Valin, 2008) there does not seem to be the same development of the remaining (activity) verbs. In RRG one of the prime motivations is the investigation of languages differing from English (Van Valin, 2008).

Case Grammar

After the initial split between state and non-state, CG variants further divide the non-state verbs into process, action and action-process (Chafe 1970). Variations split into a two dimensional matrix by adding a second parameter giving
4x4 (Cook, 1979), and 4x16 (Longacre, 1976) classes. For example, Cooke's class B.1 State Experiential Verbs contains *doubt, know, like and want.*

The CG classes offered an intermediate decomposition that would be easy to apply accurately and provided the capacity for further fine-grained decomposition. Also the non-state classes offer a clearer decomposition than those of RRG.

Consequently, the CG classes were used for the next stage of decomposition as shown in figure 3.11.

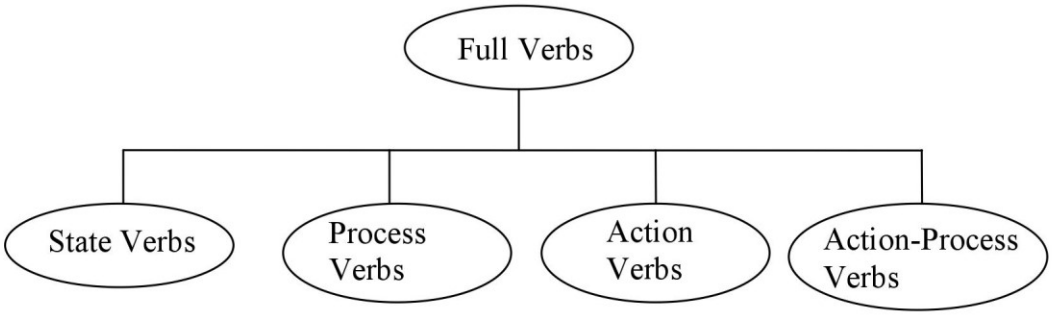


Figure 3.11 Decomposition of the Full Verbs

Levin Classes

Levin's approach (Levin, 1993) classifies verbs by alternations, methods by which verbs relate to their arguments. For example, the class of spray/load verbs participates in the Locative Alternation by taking two forms - "Sharon sprayed water on the plants." and "Sharon sprayed the plants with water." Other verbs, such as fill or pour can each take only one of the forms. Although the constraints are based on syntax, the emergent classes are believed to be semantic "… the behaviour of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large extent determined by its meaning." (Levin, 1993).

Levin's techniques use alternations to generate a tree in which there are 49 classes at the first level. 31 of these decompose to give 154 sub-classes at second level and 13 of these decompose to give 32 sub-classes at the third level. Levin's approach does not provide a good hierarchical decomposition but it does produce good classes, some of which are fine-grained. Consequently, the following decisions were made:

Three slots would be allocated to each of the CG slots, using 12 of the 14. Each of those 12 slots would also be allocated to a different Levin Class to promote semantic dispersion.

Miscellaneous verb slots and constraints

Two of the 12 slots were allocated to lower-level Levin Classes, slots 13 and 14 were each allocated to a different one of those two lower-level classes. The intention was to promote semantic proximity resulting in some higher similarity sentence pairs being generated.

It was also decided to promote polysemy by ensuring that at least one of the verbs selected appeared in several Levin categories and at least one verb appeared in only one Levin class.

As discussed under nouns, a constraint was added that one of the verbs must be part of a homonymous noun-verb pair.

As discussed under adjectives, a constraint was added that one of the verbs must be part of a homonymous verb-adjective pair. The full allocation is shown in table 3.5.

Verb classification sub-frame

Slot	CG Class	Levin Class
1	State Verb	Levin Class a
2	State Verb	Levin Class b
3	State Verb	Levin Class c
4	Process Verb	Levin Class i
5	Process Verb	Levin Class j
6	Process Verb	Levin Class k
7	Action Verb	Levin Class p
8	Action Verb	Levin Class q
9	Action Verb	Levin Class r
10	Action-Process	Levin Class x
11	Action-Process	Levin Class y
12	Action-Process	Levin Class z
13	Levin Class 2nd level	Levin Class a,b,c,i,j,k, p,q,r,x,y *OR* z
14	Levin Class 3rd level	Levin Class a,b,c,i,j,k, p,q,r,x,y *OR* z

Table 3.5 Verb classification sub-Frame

Decomposition of the adverbs

As shown in table 3.1, 8 slots in the sampling frame were reserved for adverbs. This section explains how they were allocated.

In terms of natural language processing, adverbs are probably the least studied of the four major categories of word under consideration. In the words of Jackendoff (Jackendoff, 1972)

"Most studies in generative grammar do not even concede to them the right to be a part of speech."

Jackendoff's work in generative grammar separated adverbs into categories based on the positions they can legitimately be attached to in a sentence parse tree. This approach was rejected for this work as classification could be laborious and error prone.

Modern grammar tends to classify adverbs in terms of their origin (Quirk et al., 1985) or behaviour (Carter and McCarthy, 2006) without providing much insight into their semantics. There are also some classes included in modern grammar which are adverbial forms of function words such as interrogative (e.g. when) and relative (e.g. where) and these have also been ruled out.

Consequently it was decided to use classes derived from traditional grammar. Four major categories are generally accepted: time (e.g. soon), place (e.g. here), manner (e.g. bravely) and degree (e.g. entirely). An additional class was included, frequency (e.g. often). These requirements are met by the allocation of slots 1 – 5 in table 3.6.

Miscellaneous adverb slots

Slot 6 was allocated to an adverb selected purely at random, to promote diversity of coverage.

Slots 7 and 8 were reserved in the frame for the comparison property of adverbs (comparative, superlative), discussed earlier. The final outcome of these decisions is shown in table 3.6.

Adverb classification sub-frame

Slot	Class	
1	Time	
2	Place	
3	Manner	
4	Degree	
5	Frequency	
6	Randomly selected	
7	Any of 1-6	Comparative form
8	Any of 1-6	Superlative form

Table 3.6 Adverb classification sub-Frame

3.2.1.3 Mapping the classes from the decomposition onto a sampling frame

An empty frame was constructed setting out a required combination of properties for each of the 64 word slots, which at this stage were unpopulated. So for, example, the first slot was required to be

Noun : Abstract : Quality

The first concrete noun was required to be

Concrete : Living : Biological : Body Parts

Etc.

The other constraints in the "miscellaneous" subsections of 3.2.1.2 were set out to be checked for satisfaction as the frame was filled.

3.2.1.4 Populating the slots in the frame with words using an Opportunistic Random Search

A process had to be devised to fill the slots in the frame and this was dubbed the Opportunistic Random Search. The name arises from the fact that the BNC word frequency list was searched to select words to fill the frame, that randomisation of selection was required to prevent bias, and yet a simple (opportunistic) method was used to allocate the words efficiently. The process operated as follows:

1. Divide the BNC word frequency list into two sections, the most frequent 2000 words and all the remaining lower frequency words
2. Separate the most frequent 2000 words into 4 sections: nouns, verbs, adjectives and adverbs
3. Separate the lower frequency words into 4 sections: nouns, verbs, adjectives and adverbs
4. Randomise the order of the words in each of the 8 files that were generated
5. FOR each word class (nouns, verbs, adjectives, adverbs)

    a. For the high frequency word list file, select the first word

b. IF the word meets the criteria for any high frequency slot allocate it to the slot
c. ELSE Discard the word
d. REPEAT
    i. Take the next word from the high frequency list
    ii. IF the word meets the criteria for any high frequency slot allocate it to the slot
    iii. Note any additional constraints satisfied by the word
    iv. ELSE Discard the word
e. UNTIL all the high frequency slots are filled
f. For the low frequency word list file, select the first word
g. IF the word meets the criteria for any low frequency slot allocate it to the slot
h. ELSE Discard the word
i. REPEAT
    i. Take the next word from the low frequency list
    ii. IF the word meets the criteria for any low frequency slot allocate it to the slot
    iii. Note any additional constraints satisfied by the word
    iv. ELSE Discard the word
j. UNTIL all the low frequency slots are filled

6. NEXT word class (noun, verb, adjective, adverb)

For an example, consider the allocation of high frequency verbs. Suppose the first verb selected was *know*. This is a state verb and as no verb slots would have been filled at this point, know would be allocated to the first verb slot and its Levin class recorded as 29.5 (conjecture verbs), replacing "Levin class a" in table 3.5 so that the additional criteria regarding Levin categories could be fulfilled.

Some minor modifications were required to the process to accommodate additional criteria, for example slot 6 of the adverb frame was initially allocated to any adverb occurring during the filling of the first 5 adverb frames which did not fall naturally into one of the specified categories, otherwise to be filled by purely random selection.

Slots 7 and 8 of the adverb frame were allocated to the two adverbs which most sensibly participated in the comparative/superlative variants. The interplay between the requirement to fill the slots and the satisfaction of the additional constraints made populating the table a

somewhat stochastic process, with some stages having to be re-worked. The balance between the low and higher frequency words was slightly off-target but considered acceptable. The outcome of populating the frame is shown in table 3.7 (where LF means a word is from the low frequency list).

No.	Class	Word	Additional Criteria / comments
1	Noun abstract quality	Variety	
2	Noun abstract idea from science	Temperature	
3	Noun abstract idea institutional fact	Nation	
4	Noun abstract idea	Delay	Homonymous noun-verb pair LF
5	Noun abstract state	Conflict	
6	Noun abstract emotion positive	Pleasure	
7	Noun abstract emotion negative	Doubt	
8	Noun abstract	Prejudice	Randomly selected LF
9	Noun concrete living biological body parts	Ear	
10	Noun concrete living biological fruits and vegetables	Lemon	LF
11	Noun concrete living biological animals	Cat	
12	Noun concrete living biological birds	Hen	LF
13	Noun concrete living biological insects	Fly	
14	Noun concrete living biological plants and flowers	Rose	LF
15	Noun concrete nonliving biological foodstuffs	Coffee	
16	Noun concrete nonliving nonbiological clothing	Shoe	
17	Noun concrete nonliving nonbiological tools and Manipulables	Key	Source for homophone
18	Noun concrete nonliving nonbiological vehicles	Boat	Normally found outdoors
19	Noun concrete nonliving nonbiological furniture	Chair	Normally found indoors
20	Noun concrete nonliving nonbiological musical instruments	Guitar	LF
21	Noun concrete nonliving nonbiological miscellaneous artefacts	Weapon	
22	Noun concrete nonliving nonbiological gemstones	Diamond	LF
23	Noun concrete nonliving nonbiological other non man-made	Hill	
24	Noun collective living biological	Crowd	

25	Noun collective nonliving nonbiological	Heap	LF
26	Noun concrete living biological	Parent	Randomly selected
27	Noun concrete nonliving nonbiological	Fire	Randomly selected
28	Noun Homonyn/homophone	Quay	Homophone of Key
29	Adjective dimension	Large	Source for comparative
30	Adjective physical property	Dry	Source for antonym, wet : Homonymous verb-adjective pair
31	Adjective colour	Grey	
32	Adjective age	Young	
33	Adjective Value positive	Great	NOT an antonym of terrible
34	Adjective Value negative	Terrible	NOT an antonym of great
35	Adjective speed	Slow	
36	Adjective human propensity	Responsible	
37	Adjective similarity	Familiar	
38	Adjective difficulty	Difficult	
39	Adjective qualification	Global	LF
40	Adjective smell	Rotten	LF
41	Adjective comparative	Larger	Comparative of large
42	Adjective antonym	Wet	Antonym of dry
43	Verb State (state experiential)	Hope	Levin class 32.2 long (for)
44	Verb State (state locative, continuous locative)	Cover	Levin classes 47.8 contiguous location (also 9.8 fill)
45	Verb State	Relax	Levin class 31.1 amuse LF
46	Verb Process	Change	Levin classes 41.1.1 dress (also 45.4 CoS, 26.6 turn,13.6 exchange)
47	Verb Process	Appear	Levin class 48.1 (48.1.1 appear)
48	Verb Process	Dry	Levin class 45.4 other change of state LF : Homonymous verb-adjective pair
49	Verb Action	Run	Levin classes 51.3.2 run (also 26.3 preparing, 47.5.1 swarm, 47.7 meander) Source for Levin 3rd level class pair

50	Verb Action	Laugh	Levin class 40.2 nonverbal expression
51	Verb Action	Bargain	Levin class 36.1 correspond
52	Verb Action-Process	Drink	Levin class 39.1 ingesting Source for Levin 2nd level class pair
53	Verb Action-Process	Punish	Levin class 33 judgement (negative)
54	Verb Action-Process	Delay	Levin class 53.1 lingering Homonymous noun-verb pair LF
55	Verb Action-Process	Eat	Levin class 39.1 ingesting Paired with drink
56	Verb Action	Climb	Levin class 51.3.2 run Paired with run
57	Adverb Time	Eventually	
58	Adverb Place	Far	
59	Adverb Manner	Seriously	
60	Adverb Degree	Partially	LF
61	Adverb Frequency	Mostly	LF
62	Adverb	Safely	Randomly selected
63	Adverb Comparative	Farther	
64	Adverb Superlative	Most seriously	

Table 3.7 Populated word sampling frame.

3.2.2 Experiment 3.1 Collection of the sentence pairs

Having decided upon the stimulus words, the next phase of the work was to generate 64 sentence pairs with similarity ratings. These were selected from a a database of 1,121 sentences (virtually all valid sentences), indexed by 10 different keys to support queries extracting different combinations of sentences for STSS-131 to form the 64 pairs. The difficulty in finding existing sentence datasets suggests that this database will be a useful resource in its own right for future researchers both in STSS and other fields such as Psychology and Neuroscience. The methodology for experiment 3.1 consisted of:

1. Creating a blocked design
2. Sampling the population
3. Design of the questionnaire instrument and instructions
4. Conducting a pilot study
5. Conducting the remaining trials

3.2.2.1 Creating a blocked design

A blocked design is commonly used in scientific experiments and has its origins in medicine. A typical example is where two blocks of participants suffering form an illness are selected. One block is prescribed a treatment for the illness and the other a placebo. These blocks must be balanced to control for factors which could confound experimental results such as age or gender. Comparing the results for the two blocks it is possible to determine whether the treatment is performing better than the psychological effect of receiving the placebo. The block design technique can also be used for other experimental purposes; in the following example it distributes the workload and eliminates a potential confounding factor, spurious semantic overlap. Where the experiments involve humans, the participants allocated to a particular block are often described as a group.

The original objective of the data collection experiment was to collect 1024 sentences. To achieve this, a blocked design was used in which the 64 stimulus words were divided into 4 blocks of 16. In experiment 3.1 a trial consisted of a participant writing 2 sentences for each stimulus word in a particular block. A session was an organised event at which a number of trials took place at the same time and place. The choice of number of participants is described in 3.2.2.2.

The blocked design was intended not only to provide a reasonable workload, but also to prevent spurious semantic overlap. Blocks A and B (of 32 words each) were separated so

that no pair selected from the two different blocks could have been written by the same person (or different people experiencing the same semantic context).
Blocks A and B were further subdivided (blocks A1, A2, B1, B2) to prevent internal spurious semantic overlap. For example the **adjective** *dry* appeared in block A and the **adjective** *wet* appeared in block B. However, the **verb** *dry* also appeared in block B. By splitting into blocks A1, A2, B1 and B2 it was possible to ensure that no participant experienced more than one request to produce a pair of sentences for either *wet* or *dry*.
The number of sentences collected was affected by factors such as completion failures which reduced the number of returns, but questionnaire supplements (added after piloting) acquired extra sentences.
The full block structure for experiment 3.1 is shown in appendices 3.1 and 3.2.

3.2.2.2 Sampling the population

The target sample size was set at 32 from a combination of factors such as manageable size of a trial, quantity of data expected and cost (as participants were compensated).
It was considered that the principal difficulty would be generating enough sentences so that high similarity pairs could be selected from the pool. This group size would produce (1024 x 1023)/2 = 523,776 potential sentence pairs offering scope for finding suitable similarity combinations.
The experimental task was relatively complex and demanding. It required the participants to:

- Produce valid sentences (barring errors)
- Understand the difference between nouns, adjectives, verbs and adverbs
- Demonstrate creativity in inventing sentences
- Demonstrate additional creativity in representing natural dialogue

Therefore it was decided that this experiment did not require a sample of the general population, but one which was selective in terms of creative writing ability. Therefore the target population was undergraduates on Arts and Humanities courses, who were native-English speakers and in later stages of their courses.
Recruitment of the specialised pool of graduates was achieved through a number of processes: a mailshot to lecturers, briefings at the start of classes, posters and leafleting in suitable social spaces.

All publicity materials directed the participants to a website which had details of experimental sessions to attend and also a short self-test to establish that they knew the difference between nouns, verbs, adjectives and adverbs.

Recruitment was difficult, but was helped by paying compensation for time to the participants. It was estimated that the task could be completed within one hour. The compensation was set at £5 which was a little above the minimum hourly wage in the UK at the time of the experiments.

All participants received an ethics statement / covering letter (similar to those for chapter 2) before agreeing to participate.

3.2.2.3 Design of a questionnaire instrument and instructions

There were three aspects to designing the questionnaire to capture the sentences:

- production of the basic questionnaire to capture the sentences
- incorporation of additional safeguards to prevent confusion between nouns, verbs, adjectives and adverbs
- additional features to promote accuracy of responses and the generation of high similarity pairs

Production of the basic questionnaire

Sample extracts from the questionnaire can be found in the appendices for this chapter (appendices 3.3, 3.4 and 3.5).

The basic instructions informed the participants they would be asked to produce two sentences for each stimulus word with the example:

> Please print (in BLOCK CAPITALS) two sentences, between 10 and 20 words long in the boxes below, using the Noun FIGHT

Participants were asked to write two sentences because it was anticipated that participants might quickly write the first thing that came into their heads as their initial response. For example, one participant provided the first response:

> *There is something rotten in the kingdom of Denmark.*

whereas the second response from the participant was

> *The egg had a rotten smell that hung in the room for days.*

This use of the adjective to describe a smell was exactly the kind that was desired. whereas the quote from Hamlet has very limited value as a representation of general human utterances.

Additional instructions

The participants were also given instructions on how to treat polysemous words and potential homonymous verb-noun pairs, the desirability of natural responses and the re-assurance that it was not an intelligence test. All of these instructions were intended to promote the generation of usable, natural sentences.

Each stimulus word instruction was presented on a separate page with two boxes for the responses. The pages were grouped by word type (e.g. all nouns together etc.) and an additional instruction page was presented at the start of each group.

To prevent effects of ordering, two versions of each group were produced, one with the order of the words in the group randomised, the other with that order reversed. Experience from the previously conducted experiments in this study and the nature of the task suggested that this would give sufficient protection from effects of ordering. Therefore two different questionnaires were produced for each block of words, one with all of the groups in their original random order and the other with all of the groups in their reversed order.

This instruction page contained a definition of the word type and examples of the word with preceding pronouns appropriate for the type. Adjectives and adverbs had short example phrases for their usage.

The final sheet requested minimal participant details: name, age band (to identify mature students), degree title (to confirm verbally-oriented) and a check box to confirm the participant was a native-English speaker.

Promoting high similarity - Themes

There was a concern that despite using words from related classes and even from the same, narrow class (in a few instances), there could still be a very large number of low-similarity pairs at the expense of medium and high similarity (Zipf's law, discussed in chapter 2).

A solution to this problem was found by adding an additional constraint to some of the pairs of sentences to be generated. This used an alternative approach to semantic similarity and relatedness, Thematic similarity (Klein and Murphy, 2002). Thematic similarity is relatively straightforward. Participants were asked to write a sentence on a supplied theme using the stimulus word. The assumption was that two sentences generated using the same

stimulus word and same theme were likely to have a higher similarity. Sentences using either the same word or the same theme were likely to have intermediate levels of similarity. Sentences using both different stimulus words and different themes were likely to have low similarity.

Using thematic similarity required a suitable source of themes, which were found in modern language teaching syllabuses. These focus on conversation about useful everyday activities. A set of potential themes was compiled from various examining board sources (AQA, 2010, Hawkin, 1995) and texts (Hunt, 2001, Farmer, 1996, O'Donaill and Ni Churraighin, 1995). Themes for the work in this chapter were selected on the basis of general occurrence and likelihood of being useful with the stimulus words. A full list of themes is given in appendix 3.6. Themes selected for use were:

- *Going out (socially), giving invitations*
- *Future Events, speaking about the future*

Each of the questionnaires had a theme applied to the final noun presented to the participants. An extra sheet explaining the task was inserted before the final noun, and an example of the instructions on the noun page is:

> Please print (in BLOCK CAPITALS) two sentences, between 10 and 20 words long in the boxes below, on the general Topic of
>
> *Going out (socially), giving invitations*
>
> and using the Noun VARIETY

3.2.2.4 Conducting the pilot study

There were a number of unknown factors that could affect the operation of the trials to collect the sentences. How demanding was the task? It was planned to use one hour time slots to perform the trials. Would the task prove too demanding (the trials might overrun) or too simple (the trials might finish early)? How attractive was the compensation? Would the trials be over-subscribed or would it be impossible to attract enough participants? What would the quality of the participant responses be under the specific controlled circumstances of the trials?

It was considered that procedural issues could be checked by reviewing the result of the first 8 participants recruited. If there were no obvious problems, these would be included in

the full study. The participants were recruited from undergraduates in the languages department. The following observations were made:

- The participants were able to complete the tasks without difficulty.
- The participants were able to complete the tasks with time to spare.
- The participants found the compensation acceptable, but were not motivated purely by the financial element.
- The participants had tended to rush through the task, the environment seemed to engender a feeling of trying to finish an exam quickly.
- The participants had a strong tendency to respond with assertions but not other dialogue acts (questions, instructions, expressions etc.).
- Concerns about the capability of generating higher similarity sentence pairs appeared to have some grounds.

3.2.2.5 Conducting the experimental trials

In the light of the pilot study the following decisions were made:

- to include the 8 trials from the pilot in the final study
- to modify the instructions to promote a wider range of dialogue acts
- to modify the instructions to inform participants at the start that they would be able to do more if they finished early
- to produce questionnaire supplements providing additional work for participants who finished early
- to include additional themes to promote the likelihood of high and medium-high similarity combinations in the supplements
- where circumstances permitted, to play ambient background music at a low level to reduce the feeling of being in an exam
- to canvas more extensively and with more effort to recruit students to further trials

It should be emphasised that the materials produced by the pilot study were perfectly good examples of the kinds of sentences required, apart from the facts that they appeared to be less likely to produce higher similarity combinations at this stage and that they did not cover the diverse range of dialogue acts desired. However, it was felt that they would be useful when combined with those produced by the remaining trials.

To promote a wider range of dialogue acts, the first page of the questionnaires was modified to contain the instruction:

> Your sentences should be in the form of statements, instructions, commitments, expressions or declarations.

An additional sheet was also provided to the participants, clarifying the difference between statements, instructions, commitments, expressions or declarations. This is shown in appendix 3.7.

The themes extracted from the language teaching sources described earlier were used in two ways:

- to encourage the formation of high-similarity pairs when using the same stimulus word
- to encourage the formation of medium-high similarity pairs by applying a theme to a stimulus word and its synonym (derived from a thesaurus for a small subset of the stimulus words)

The additional themes are shown in appendix 3.8.

There were continuing difficulties in recruiting students due to the approaching exam period, but 29 from the target of 32 participants were recruited and the use of the supplements meant that 1,121 sentences were collected – in excess of the original target of 1024. Also good coverage of the blocks in the design was achieved, with 7 questionnaires returned for each block and one extra questionnaire for block A1.

The outcome of providing the examples of word classes in the instructions was that word class errors were limited to approximately 1.6% of the sentences.

The students in general took the task seriously and expressed an interest in the work. However, two of them either did not take the process seriously or had an unusual world view dominated by fantasy gaming, one extreme case being:

> "It was so large!" exclaimed Hrothgar the barbarian, "that large sword, it was truly epic!"

These sentences were not used. A sample of the sentences obtained from the process may be found in appendix 3.9.

3.2.3 Selection of the 64 sentence pairs

It is clear from the work in chapters 1 and 2 that (barring the algorithms we wish to test with STSS-131) the only way to determine the similarity of a pair of sentences is through human judgement. Therefore the only way to select pairs of sentences from the pool was through human judgement (guided by the language representation requirements discussed when creating the sampling frame in section 3.2.1.2).

Selecting the 64 sentence pairs for the dataset from the pool of 1,125 candidate sentences required four stages of work to be completed:

- creation of a database with report generation facilities for a subset of the collected sentences
- convening a committee of 3 judges to select 64 pairs of sentences for the dataset
- conducting a short pilot study
- revising the set of 64 sentences and the procedures in the light of the pilot study

3.2.3.1 Creation of the database

A database was created to act as a repository for the data. Individual tables were created for participants, sentences, and each of the 4 word classes – nouns, verbs, adjectives and adverbs. The participants table contained the basic identification information, age band, course, questionnaire and supplement (so the variant of the questionnaire and supplement identified by each participant could be traced).

The sentence table contained the participant number, stimulus word, stimulus word class, stimulus theme, speech act, and comment information. Each of the different stimulus word tables contained a set of fields appropriate to its type. For example, the nouns table contained the noun itself, type (abstract, concrete etc.), Booleans for living and for biological, class (e.g. CSD category) and a number of fields for its relationship to other stimulus words in the set (e.g. the noun delay has a homonymous verb-noun pair relationship with the verb delay).

Over 170 queries were written to extract groups of sentences from the database. A priori knowledge was used to set criteria to produce collections that were likely to provide various degrees of similarity. These included queries that would return all sentences generated for a particular stimulus word, all sentences for each of a pair of stimulus words (e.g. cat and hen which have the common features Noun:concrete:living:biological) and all sentences generated by a particular theme or combination of themes (e.g. family and friends combined with qualities of a good friend).

Queries were also written to select sentences for a set of pairings between the words in blocks A and B, partly selected to promote various degrees of similarity and partly to ensure that every stimulus word had the opportunity to appear at least once in the final dataset.

3.2.3.2 Materials selection by committee

Each member of the committee was supplied with the printed queries and asked to search for combinations of sentences that were, in the view of the judge, likely to be rated as of high, medium or low similarity by human raters. Two of the judges were highly experienced in scripting dialogue for conversational agents and the third was highly experienced in interviewing users of conversational agents as part of the knowledge engineering process.

There were many areas of disagreement where the views of the judges had to be examined thoroughly before consensus was reached. Because of the difficulty involved in reaching consensus, it was decided to pilot the 64 sentence pairs before conducting the full experimental program to collect the ratings.

At this stage it was also planned to dedicate 4 of the sentence pairs to an investigation of dialogue acts, as it was felt that factor analysis could provide some insight into their influence on perceived semantic similarity. This decision is discussed further in section 3.2.4.1.

Two more sentence pairs were added to bring the set up to 66. These were calibration pairs. One was the sentence pair with the lowest consistent similarity ratings and the other with the highest set of similarity ratings from STSS-65. There were two benefits from this. First it validated STSS-65 by determining whether or not the users had expanded the similarity range (which could have happened because all of the dialogue acts were assertions). If this were the case one would expect that they would be closer in value when mixed in with STSS-131 sentences. Second they ensured that the range of similarities spanned by STSS-131 was at least as large as the range covered by STSS-65.

3.2.4 Experiment 3.2 Collecting similarity ratings for the sentence pairs

Due to the difficulties the committee experienced about reaching consensus on the selection of sentence pairs, a pilot study was conducted to investigate whether the set of 64 new sentence pairs was likely to have a good representation of the range of similarities when rated by human participants. In experiment 3.2 a trial consisted of a participant rating

the semantic similarity of 64 sentence pairs. A session was an organised event at which a number of trials took place at the same time and place.

3.2.4.1 Pilot study and modifications to materials

The target population was 16 undergraduate students. This study was intended to validate the sentence pairs and it was considered unlikely that the results would be incorporated into a dataset. As a result, it was not sensitive to population sampling. For convenience and speed, computing students were used who volunteered without compensation. Benefits for the participants included furthering their knowledge of AI concepts that could support education modules taken; also gaining firsthand experience of questionnaire-based evaluation techniques that could be applied to their own final year projects.

Fourteen students were recruited within the available time for the pilot study, 9 males, 2 females and 1 withheld personal data. Eight of the participants were aged 18-22 and three were older than 22. The students were drawn from 7 different courses ranging from soft (verbally-oriented) to hard (numerically/symbolically-oriented) computing.

The sentence pairs were numbered to simplify processing, but the allocation of the sentence pairs to the numbers was randomised to prevent any inference being made about degree of similarity from the card number.

Participants were allowed to complete the process in their own time, subject to the usual instructions about completing in a single sitting without distraction etc.

The first outcome of the pilot study was the discovery that the 4 sentence pairs for dialogue acts were not producing enough data for an ANOVA analysis so it was decided to drop this aspect from STSS-131.

Once the 2 calibration pairs and the 4 pairs intended for dialogue acts were removed, the distribution of the remaining pairs was as shown in table 3.8.

Similarity Range	Number of Sentence Pairs
0.00 – 0.99	27
1.00 – 1.99	13
2.00 – 2.99	13
3.00 – 4.00	7

Table 3.8 The distribution of similarity ratings in new dataset pilot study

The results of the experiment showed that the judges had been optimistic in their selection of candidate pairs for high similarity. Consequently some of the sentence pairs were replaced with others more likely to achieve high similarity.

Modifying the materials was not as simple as, say, replacing 7 of the pairs from the low similarity band. By this stage there was a great deal of interdependency between the pairs in satisfying the various combinations of representation criteria.

The 4 sentence pairs originally intended to investigate dialogue acts were released for use in meeting the high similarity requirement. A further 7 slots were identified, 5 of which had been predicted to achieve high or medium similarity and failed to do so, and two of which contained low frequency pairs whose contribution to the low-frequency sample could be fulfilled by other sentence pairs.

The replacement materials were generated by taking 11 sentences from the dataset and asking a group of 16 graduate participants to paraphrase them. Each participant paraphrased 3 sentences, providing between 3 and 6 options to combine with each original sentence. The graduate participants (who had English teaching experience) were recruited using the same process as for STSS-65.

The technique was chosen because it was likely to produce some usable high similarity pairs (which did not necessarily use the same words or word orders). This was the final option to obtain a suitable balance of materials before expending significant effort and expense in obtaining the ratings, so a high degree of confidence was needed. The sentences selected for paraphrase generation are listed in appendix 3.10, an example paraphrase sentence capture sheet is given in appendix 3.11 and a sample of paraphrased sentences is listed in appendix 3.12

The final set of sentence pairs used is given in appendix 3.13. The cards were produced to the same specification as the card sort with semantic anchors experiment described in chapter 2.

3.2.4.2 Specification, selection and recruitment of a suitable population sample

As experiment 3.2 was the single most important data collection experiment in the whole of the work in this thesis considerable effort went into acquiring a good population sample. A blocked design was used with 2 groups, each with target sizes of 32 participants. The first group consisted of volunteers from the general population who volunteered without compensation; the second consisted of undergraduate students who were compensated at a rate of £5. In all cases the task was timed to last about an hour.

There were three reasons for the choice of sample groups and their sizes.

1. It would allow a comparison of the ratings returned by the two groups to search for evidence as to whether a student sample is representative of the population as a whole.

2. The block size of 32 would allow each sample to be considered statistically significant if they were considered to come from different populations (based on experience in chapter 2).
3. If statistical tests provided evidence that the blocks come from the same population, they could be combined to produce an even more persuasive set of results.

Whether or not a sample of students represents the general population is an issue which has not been investigated throughout nearly half a century of research on word similarity. Collecting evidence about this should be very useful to future researchers in the field.

Demographic principles

The purpose of controlling the demographics was to ensure that there were no confounding factors. Those considered to be important were age, gender and academic background. Within the groups (students, non-students) gender should be evenly balanced between male and female and academic background should be balanced between science/engineering vs arts/humanities. Within each demographic group the age distribution of that element of the population should be represented, with most of the students being in the 18-22 band and most of the non-students being in the over 21 bands.

Between the groups, the age factor should be different, as explained above, whereas gender and academic background should have similar distributions.

3.2.4.3 Experimental trials

The undergraduates completed the task at one of a number of supervised sessions organised in their faculties. General population volunteers completed in their own time following the usual instructions to set aside an hour when they would be undisturbed to focus exclusively on the work.

Previous experience with the experiments in chapter 2 provided evidence that the card sorting task was sufficiently clear to be performed by members of the general population without supervision and no problems were reported by the participants. Running supervised sessions for the students was however, more administratively convenient and it eliminated problems of non-returns or late returns from that sample group. Once again, the compensation was adequate but not excessively attractive, so significant effort (canvassing, leafleting, posters etc.) was required to recruit student participants and a number of sessions had to be run over a period of two months.

The outcome of these sessions was that the targets were met and slightly exceeded, so the first 32 returns were used for each group to avoid bias.

3.2.4.4 Results of discriminability experiment

The results of the data trials are shown in table 3.9, with the actual sentence pairs. The first column, SP, is the identification number of the sentence pair, continuing the numbering sequence from the sentences pairs in Chapter 2.

Column 2 lists the two sentences comprising the pair. Column 3 is the mean of the student ratings for each sentence pair (n=32) and column 4 is the mean of the general population ratings for each sentence pair (n=32). Column 5 is the mean of participants in the combined student and general population groups (n=64).

SP	Sentences	Student Rating	Non-student	Combined Rating
66	Would you like to go out to drink with me tonight? I really don't know what to eat tonight so I might go out somewhere.	0.97	1.05	1.01
67	I advise you to treat this matter very seriously as it is vital. You must take this most seriously, it will affect you.	3.12	3.63	3.38
68	When I was going out to meet my friends there was a delay at the train station. The train operator announced to the passengers that the train would be delayed.	3.17	3.09	3.13
69	Does music help you to relax, or does it distract you too much? Does this sponge look wet or dry to you?	0.08	0.11	0.1
70	You must realise that you will definitely be punished if you play with the alarm. He will be harshly punished for setting the fire alarm off.	3.05	2.63	2.84
71	I will make you laugh so much that your sides ache. When I tell you this you will split your sides laughing.	3.66	3.85	3.75
72	You shouldn't be covering what you really feel. There is no point in covering up what you said, we all know.	2.09	2.33	2.21
73	Do you want to come with us to the pub behind the hill? We are going out for drinks tonight in Salford Quays if you would like to come.	1.97	1.67	1.82
74	This key doesn't seem to be working, could you give me another? I dislike the word quay, it confuses me, I always think of things for locks, there's another one.	0.88	0.57	0.72
75	The ghost appeared from nowhere and frightened the old man. The ghost of Queen Victoria appears to me every night, I don't know why, I don't even like the royals.	1.37	1.53	1.45
76	You're not a good friend if you're not prepared to be present when I need you. A good friend always seems to be present when you need them.	2.94	3.33	3.14
77	The children crossed the road very safely thanks to the help of the lollipop lady. It was feared that the child might not recover, because he was seriously ill.	0.19	0.07	0.13

SP	Sentences	Student Rating	Non-student Rating	Combined Rating
78	I have invited a variety of people to my party so it should be interesting. A number of invitations were given out to a variety of people inviting them down the pub.	2.43	1.93	2.18
79	I offer my condolences to the parents of John Smith, who was unfortunately murdered. I express my sympathy to John Smith's parents following his murder.	3.87	3.95	3.91
80	Boats come in all shapes and sizes but they all do the same thing. Chairs can be comfy and not comfy, depending on the chair.	0.59	0.41	0.5
81	If you continuously use these products, I guarantee you will look very young. I assure you that, by using these products consistently over a long period of time, you will appear really young.	3.62	3.54	3.58
82	We ran farther than the other children that day. You ran farther than anyone today.	2.29	2.57	2.43
83	I always like to have a slice of lemon in my drink especially if it's Coke. I like to put a wedge of lemon in my drinks, especially cola.	3.74	3.88	3.81
84	It seems like I've got eczema on my ear doctor, can you recommend something for me? I had to go to a chemist for a special rash cream for my ear.	2.21	1.89	2.05
85	I am proud of our nation, well, most of it. I think of myself as being part of a nation.	1.7	1.73	1.71
86	There was a heap of rubble left by the builders outside my house this morning. Sometimes in a large crowd accidents may happen, which can cause deadly injuries.	0.12	0.06	0.09
87	Water freezes at a certain temperature, which is zero degrees Celsius. The temperature of boiling water is 100 C and the temperature of ice is 0 C.	2.9	3.25	3.08
88	We got home safely in the end, although it was a long journey. Though it took many hours travel, we finally reached our house safely.	3.48	2.65	3.06
89	A man called Dave gave his fiancée a large diamond ring for their engagement. The man presented a diamond to the woman and asked her to marry him.	3.25	3.19	3.22

SP	Sentences	Student Rating	Non-student	Combined Rating
90	I used to run quite a lot, in fact once I ran for North Tyneside. I used to climb lots at school as we had a new climbing wall put in the gym.	0.87	0.61	0.74
91	I love to laugh as it makes me happy as well as those around me. I thought we bargained that it would only cost me a pound.	0.07	0.09	0.08
92	Because I am the eldest one I should be more responsible. Just because of my age, people shouldn't think I'm a responsible adult, but they do?	2.15	2.31	2.23
93	I need to dash into the kitchen because I think my chip pan is on fire. In the event of a chip pan fire follow the instructions on the safety note.	1.62	1.79	1.7
94	Peter was a very large youth, whose size intimidated most people, much to his delight. Now I wouldn't say he was fat, but I'd certainly say he was one of the larger boys.	2.21	1.72	1.96
95	I'm going to buy a grey jumper today, in half an hour. That's a nice grey top, where did you get it from?	1.5	1	1.25
96	We got soaked in the rain today, but now we are nice and dry. I was absolutely soaking wet last night, I drove my bike through the worst weather.	1.72	1.64	1.68
97	Global warming is what everyone is worrying about today. The problem of global warming is a concern to every country in the world at the moment.	3.31	2.96	3.14
98	He was harshly punished for setting the fire alarms off. He delayed his response, in order to create a tense atmosphere.	0.36	0.09	0.22
99	Midday is 12 o'clock in the middle of the day. Noon is 12 o'clock in the middle of the day.	3.92	4	3.96
100	That's not a very good car, on the other hand mine is great. This is a terrible noise level for a new car.	1.41	0.69	1.05
101	There was a terrible accident, a pileup, on the M16 today. It was a terrible accident, no one believed it was possible.	2.2	2.47	2.33

Sentence Pair	Sentences	Student Rating	Non-student	Combined Rating
102	After hours of getting lost we eventually arrived at the hotel. After walking against the strong wind for hours he finally returned home safely.	1.4	0.78	1.09
103	The first thing I do in a morning is make myself a cup of coffee. The first thing I do in the morning is have a cup of coffee.	3.78	3.92	3.85
104	Someone spilt a drink accidentally on my shirt, so I changed it. It appears to have shrunk, it wasn't that size before I washed it.	0.6	0.36	0.48
105	I'm worried most seriously about the presentation, not the essay. It is mostly very difficult to gain full marks in today's exam.	1.04	0.5	0.77
106	It is mostly very difficult to gain full marks in today's exam. The exam was really difficult, I've got no idea if I'm going to pass.	2.29	2.8	2.54
107	Meet me on the hill behind the church in half an hour. Join me on the hill at the back of the church in thirty minutes time.	3.92	3.94	3.93
108	If you don't console with a friend, there is a chance you may hurt their feelings. One of the qualities of a good friend is the ability to console.	2.8	3.23	3.01
109	We tried to bargain with him but it made no difference, he still didn't change his mind. I tried bargaining with him, but he just wouldn't listen.	3.42	3.43	3.43
110	It gives me great pleasure to announce the winner of this year's beauty pageant. It's a real pleasure to tell you who has won our annual beauty parade.	3.84	3.91	3.88
111	They said they were hoping to go to America on holiday. I like to cover myself up in lots of layers, I don't like the cold.	0.25	0.06	0.16
112	Will I have to drive far to get to the nearest petrol station? Is it much farther for me to drive to the next gas station?	3.73	3.95	3.84
113	I think I know her from somewhere because she has a familiar face. You have a very familiar face, where do I know you from?	3.26	3.45	3.36

Sentence Pair	Sentences	Student Rating	Non-Student	Combined Rating
114	I am sorry but I can't go out as I have a heap of work to do. I've a heap of things to finish so I can't go out I'm afraid.	3.71	3.48	3.6
115	The responsible man felt very guilty when he crashed into the back of someone's car. A slow driver can be annoying even though they are driving safely.	1.22	0.53	0.88
116	Get that wet dog off my brand new white sofa. Make that wet hound get off my white couch – I only just bought it.	3.74	3.44	3.59
117	He fought in the war in Iraq before being killed in a car crash. The prejudice I suffered whilst on holiday in Iraq was quite alarming.	0.45	0.65	0.55
118	The cat was hungry so he went into the back garden to find lunch. The hen walked about in the yard eating tasty grain.	1.05	1.35	1.2
119	My bedroom wall is lemon coloured but my mother says it is yellow. Roses can be different colours, it has to be said red is the best though.	0.54	0.83	0.68
120	Would you like to drink this wine with your meal? Will you drink a glass of wine while you eat?	3.61	3.51	3.56
121	Roses can be different colours, it has to be said red is the best though. Roses come in many varieties and colours , but yellow is my favourite.	2.43	3.23	2.83
122	Flies can also carry a lot of disease and cause maggots. I dry my hair after I wash it or I will get ill.	0.13	0.11	0.12
123	Could you climb up the tree and save my cat from jumping please? Can you get up that tree and rescue my cat otherwise it might jump?	3.77	3.89	3.83
124	The pleasure that I get from studying, is that I learn new things. I have a doubt about this exam, we never got to study for it.	0.68	0.8	0.74
125	The perpetrators of war crimes are rotten to the core. There are many global issues that everybody should be aware of, such as the threat of terrorism.	1.28	0.63	0.95
126	The damp was mostly in the very corner of the room. The young lady was somewhat partially burnt from the sun.	0.1	0.11	0.11

Sentence Pair	Sentences	Student Rating	Non-student	Combined Rating
127	We often ran to school because we were always late. I knew I was late for my class so I ran all the way to school.	2.73	3.47	3.1
128	I hope you're taking this seriously, if not you can get out of here. The difficult course meant that only the strong would survive.	0.76	0.23	0.05
129	The shores or shore of a sea, lake or wide river is the land along the edge of it. An autograph is the signature of someone famous which is specially written for a fan to keep.	0.08	0.14	0.11
130	I bought a new guitar today, do you like it? The weapon choice reflects the personality of the carrier.	0.23	0.09	0.16
131	I am so hungry I could eat a whole horse plus dessert. I could have eaten another meal, I'm still starving.	3.13	2.98	3.06

Table 3.9 Similarity ratings for STSS-131 for students and non-students

Ratings in table 3.9 have been rounded to 2 decimal places as this is the greatest precision that can be obtained under the interpolated digit assumption used in standard physical science experiments. In all cases rounding was applied after the means had been calculated from the raw ratings.
Calculating the Pearson product-moment correlation coefficient for the unrounded data gives 0.971672 and calculating it for the data rounded to 2 decimal places gives 0.97. The correlation coefficients include the two calibration pairs (SP99 and SP129) carried over from the experiments in chapter 2. This is appropriate in calculations to determine whether the two groups can be sensibly combined, but SP99 and SP129 should not be included when the dataset is being used to evaluate STSS algorithms.

3.2.4.5 Analysis of results

The demographic targets for the population sample were achieved.

In the case of age, although there were no non-students in the 51-60 band, there were participants in the bands on either side. The desired distinctiveness in age distributions was achieved. The results for the age distribution are shown in table 3.10

Students	
Age band	Number
< 18	0
18-22	27
>22	2
Withheld	3

Non-students	
Age band	Number
<21	0
21-30	9
31-40	7
41-50	6
51-60	0
>60	3
Withheld	7

Table 3.10 Age distributions for Student and Non-Student population samples

As for gender, a good balance between male and female was obtained from both groups (for those who supplied their gender) and a similar number withheld in each group. The results for the gender distribution are shown in table 3.11.

Gender	Number in Students	Number in non-students
M	12	14
F	13	13
Withheld	7	5

Table 3.11 Gender distributions for Student and Non-Student population samples

For the students, current degree topic was recorded and classified into one of the two categories. Some degrees straddle the arts/humanities divide (e.g. Psychology) and these are recorded as Interdisciplinary. For non-students the highest educational qualification was recorded. The graduates were almost perfectly balanced between arts and sciences. Of the remainder, no qualification, GCSE and professional can probably be considered to occupy the middle ground, whereas the HND and MSc are from the Science/Engineering category. Therefore between groups there was an almost identical slight bias towards science/engineering. The balance both within and between groups was obtained, as shown in table 3.12.

Students	
Degree discipline	Number
Arts/Humanities	12
Science/Engineering	15
Interdisciplinary	2
Withheld	3

Non-students	
Highest qualification	Number
None	4
GCE/GCSE	3
HND (Engineering)	1
BSc	7
BA	8
MSc	2
Professional Qualification	3
Withheld	4

Table 3.12 Educational background distributions for Student and Non-student samples

It should be noted that 3 of 4 declaring no qualification in the non-students were also over 60. There was a high proportion of university graduates (53%) among the non-students. This can be explained, at least in part, by the nature of the region which has a strong tradition of engineering and technology industries, with more recently developed IT, media and digital cultural industries.

The steps taken after the pilot data collection study to increase the number of high similarity sentence pairs were effective, as shown in figure 3.12. This histogram represents the 64 new sentence pairs in the dataset (calibration pairs 99 and 129 omitted). It also uses the combined ratings from both student and non-student samples (n=64). It shows that the increase in high similarity pairs was obtained somewhat at the expense of the medium similarity bands (1.0 – 2.99). If the low-medium and medium-high bands are considered as a single, larger, medium similarity band however, the breakdown becomes 20 low, 20 medium and 24 high.

Clearly one area for future work is the collection of more medium-similarity sentence pairs, nevertheless as the most common judgement required in Conversational Agents and text mining is the distinction between high similarity and the other degrees of similarity, the composition of this benchmark data set is appropriate.

The combination of the student and non-students samples begs the questions of whether this pooling of data is permissible or useful. Rubenstein & Goodenough (Rubenstein and Goodenough, 1965) set a precedent by combining population samples from word similarity rating experiments on the basis of a high correlation between groups.

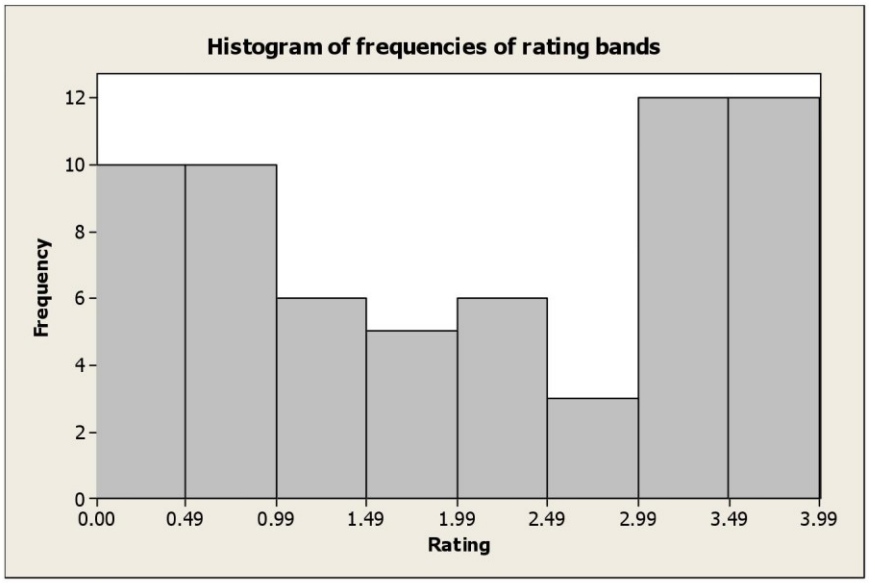


Figure 3.12 Distribution of similarity ratings in new dataset

They combined two samples (n=15 and n=36) from the same candidate population of students on the basis that the Pearson product-moment correlation between them was 0.99. It therefore seems quite reasonable to combine the two groups (n=32, n=32) from the experiment in this section given their correlation coefficient of 0.97 and the hypothesis that they are *different* kinds of samples representing the *same* population.

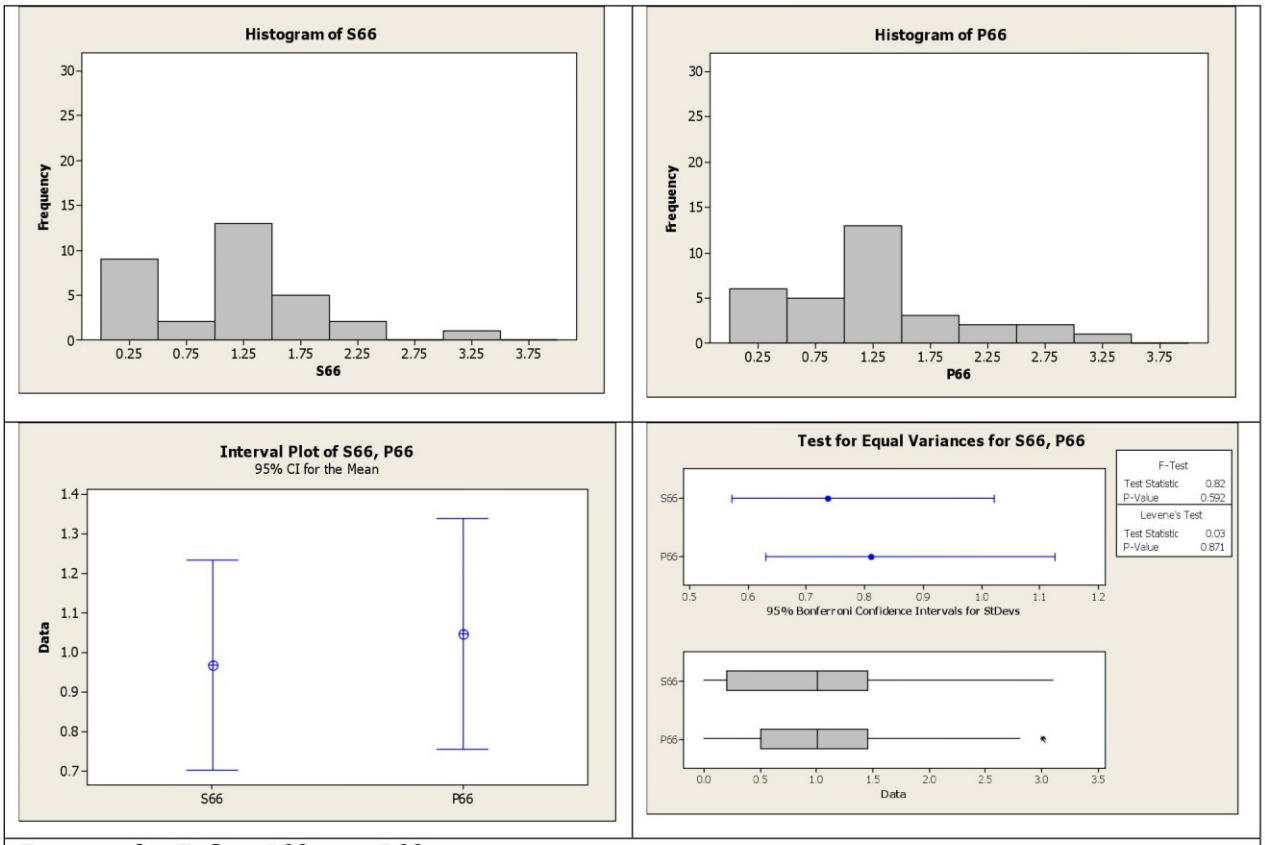
The other aspect of combining the two samples is value gained by doing so.

A substantial investigation was undertaken, using Minitab to answer the questions:

- Can the student and non-student samples be legitimately combined into a single (n=64) sample which better represents the general population?
- What additional information can be gained by combining the student and non-student samples?

Can the student and non-student samples be legitimately combined?

For each of the 66 sentence pairs (including the calibration set), histograms were plotted for the student and non-student responses, an interval plot was produced to compare the means of the ratings for the students and non-students and two tests for equal variances were conducted (using Minitab). As an example, the materials for sentence pair 66 are shown in figure 3.13:




```
Two-sample T for S66 vs. P66

      N   Mean  StDev  SE Mean
S66  32  0.969  0.735     0.13
P66  32  1.047  0.810     0.14


Difference = mu (S66) - mu (P66)
Estimate for difference:  -0.078125
95% CI for difference:  (-0.464643, 0.308393)
T-Test of difference = 0 (vs not =): T-Value = -0.40  P-Value = 0.688  DF = 62
Both use Pooled StDev = 0.7734
```

Figure 3.13 Comparison of Student and Non-student ratings for sentence pair 66

The interesting features are

- the modes of the two histograms are identical, but the distribution about the mode varies
- the interval plot shows that the means are close in value and there is a large overlap in the 95% confidence intervals indicating visually that the means are not significantly different
- the F-Test had a p-value of 0.592 and Levene's test has a p-value of 0.871 indicating that the two samples are from populations with equal variances.
- the two extremes of the 95% confidence intervals (-0.464643, 0.308393) include zero, which suggests that there is no significant difference between the means

To represent the same population, the two samples should have the same variance. The F-test and Levene's test are both well-known tests for equal variance. The F-test assumes a normal distribution and Levene's test assumes the data comes from a continuous, but not necessarily normal, distribution. Both statistics have been quoted due to issues discussed in chapter 2. For both the tests, the null hypothesis is that the variances of the two populations are equal. Normal significance values (α, the chance of finding an effect that does not exist) are 0.05 and 0.01. If the p-value is greater than α (which is clearly the case), the null hypothesis should not be rejected.

What additional information can be gained by combining the student and non-student samples?

First, increasing *n* from 32 to 64 increases the statistical significance of the sample and this has value in itself (although previous experiments in chapter 2 provided evidence that n=32 is sufficient for good quality ratings).

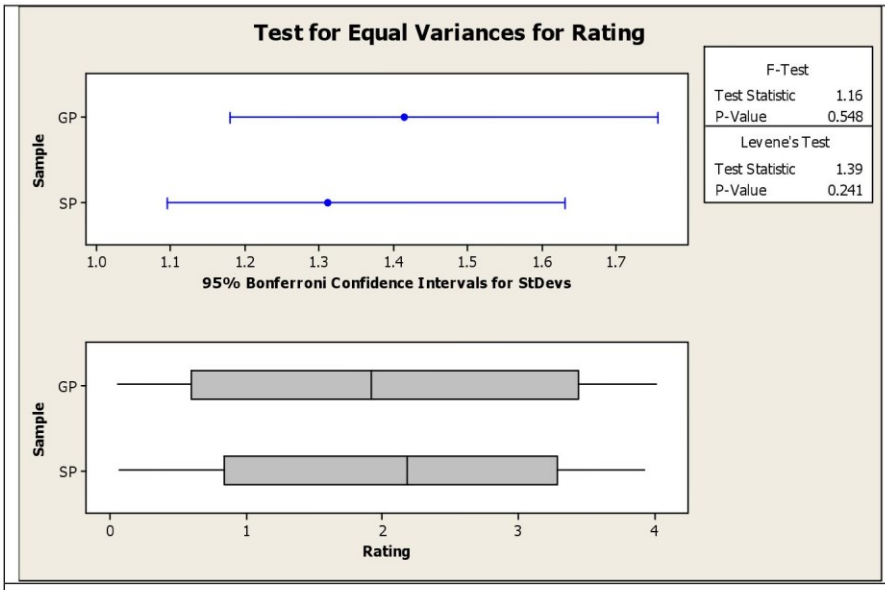
However, another question is can we improve upon the precision by increasing the number of ratings? This is another occurrence of the difficult problem of not knowing the ground truth. However, it is possible to make the assumption that if the samples pass the variance test, and if ratings produced by the individual samples (students and non-students) are not too radically different, the result of combining them will be closer to the ground truth for the overall population.

Taking sentence pair 66 as an example, again, the ratings were: students 0.97, non-students 1.05 and a combined rating of 1.01. Little benefit was likely to be gained by combining the samples in this case because the confidence interval data indicated no significant difference between the means.

On the other hand, sentence pair 104 had ratings of 1.41 from students, 0.69 from non-students and a combined rating of 1.05. It passed all tests for equality of variances but the interval test indicated significantly different means, consequently this pair benefited from the production of a genuinely different rating from the combination process.

Now it would be unscientific to pick and choose pairs to combine from the overall set, either all must be combined or none. Looking at proportions of individual cases may be interesting but still does not reveal the full picture. A much more informative view was obtained by using the General Linear Model in Minitab to combine information from all of the sentence pairs in a single judgement (results shown in figure 3.14).

The outcome of the complete GLM test was as follows:




```
95% Bonferroni confidence intervals for standard deviations

Sample   N    Lower    StDev    Upper
    GP  66  1.18118  1.41430  1.75605
    SP  66  1.09600  1.31230  1.62941


F-Test (normal distribution)
Test statistic = 1.16, p-value = 0.548


Levene's Test (any continuous distribution)
Test statistic = 1.39, p-value = 0.241
```

Figure 3.14 Test for equal variances in Student and Non-student samples

Interpretation

The p-values of 0.548 and 0.241 are greater than reasonable choices of α, so the null hypothesis of the variances being equal should not be rejected. This data does not provide enough evidence to claim that the populations have unequal variances. Therefore the two samples can be combined to provide a larger, more statistically powerful sample.
The outcome of combining the samples was determined by reviewing the confidence interval data, which showed that 19 of the ratings would be modified significantly (improved) by combination procedure.

Calibration

Two sentence pairs, SP99 and SP129, were included in the new dataset because they were the highest and lowest rated pairs in the original dataset. There were two reasons for this. Firstly, to ensure that the materials in the new dataset spanned a range of semantic

similarity values at least as wide as the original dataset. Second, it provided a way of investigating whether the human ratings were influenced by context.

This is because in the original dataset all sentences were statements – dictionary definitions of the stimulus words, whereas in the new dataset a diverse range of speech acts and stimulus words was involved. Therefore it was possible that amongst the new dataset the ratings might change if context was a factor – more specifically it was possible that the low similarity pair, SP129, might be rated higher due to the common speech act and purpose between the two sentences. The results of the ratings for these sentence pairs are reproduced in table 3.13

Sentence pair	Original dataset (chapter 2)	New Dataset (chapter 3)
SP129/SP5	0.02	0.11
SP99/SP64	3.82	3.96

Table 3.13 Calibration test

The results show that there is a change, but not a substantial one between the ratings in the two different contexts. A two sample t-test conducted for the low similarity pair with summarised data in Minitab gave a 95% confidence interval range of (-0.2017, 0.0217) and a P-Value = 0.113. A two-sample t-test for the high similarity sentence pair gave a 95% confidence interval range of (-0.3384, 0.0584) and a P-Value = 0.161. In both cases the CI range includes zero, indicating no significant difference between the means and, since the p-value is greater than commonly chosen α-levels, that also suggests no evidence for a difference between the values for SP95 and SP5.

Overall there is slight evidence for the hypothesis that the sentence pairs will appear more similar in a more diverse context; however the 95% CIs and p-values provide much stronger evidence for the overall robustness of the measurement technique to contextual influences.

Inter-rater agreement

Table 3.14 compares the relevant correlations for STSS-65 and STSS-131. Noise is the sum of the standard deviations of all sentence pair ratings over the relevant dataset (as calculated in chapter 2).

Best, mean and worst human correlations (calculated for a particular participant and all of the rest of the participants in each case) establish performance baselines for an STSS algorithm.

Dataset	Mean r	Best participant r	Worst participant r	Noise
STSS-131	0.891	0.951	0.678	0.696
STSS-65	0.938	0.976	0.830	0.360

Table 3.14 Inter-rater agreement for both benchmark datasets

A realistic target is for the algorithm to perform as well as the average human and a measure which approached the correlation of the best human would be doing very well. Comparison of STSS-131 with STSS-65 suggests that it is more demanding on the human raters, as they have not performed to the same level. This is in terms of inter-rater agreement and also the noisiness of the data, as measured by the mean of the standard deviations of the ratings over the set of sentences.

The main explanation for the lower inter-rater agreement in this chapter's work is likely to be the greater diversities of both the human population sample and the English language sentence population sample. The work in this chapter uses a mixture of students and non-students from backgrounds and ages (more of the younger participants). The sentences used are derived from much more diverse stimulus words and encompass a greater range of dialogue acts. It should not be surprising if STSS algorithms also find this dataset more challenging as that is an intended outcome. Lower scores amongst current algorithms will allow scope for measurement of improvement in future algorithms.

3.2.4.6 Application of STSS-131: Comparing the performance of two different STSS measures

This section illustrates the use of the new dataset, created in this chapter, to compare two STSS measures – STASIS and LSA.

The value of a semantic similarity algorithm is measured by its correlation with human judgement. This is because STSS-based CAs will be interested in the degree of matching (i.e. semantic similarity) between a user utterance and utterances from a set stored within the agent.

- High similarity with one of the stored utterances implies that the rule storing that utterance should be activated.
- Failing this, medium similarity between the user utterance and one of the stored utterances suggests that the user may have partial understanding or is presenting an unanticipated variant of one of the cases dealt with by the agent, which needs to be explored through questioning.

- Low similarity with all of the stored utterances suggests that the user has gone off topic and needs to be redirected.

A high correlation between the agent and human judgements over a benchmark dataset implies that the agent will make sensible judgements about user utterances.
In this comparison we would expect an STSS measure to equal or exceed the mean of the inter-rater agreements, which is a correlation of 0.891. Anything approaching 0.951would be considered to be highly successful. The worst human performance of 0.678 sets the lower bound for performance. The inter-rater agreement between the two samples (students and non-students) is even higher at 0.970 so performances of up to this figure by algorithms will still be credible.
Table 3.15 shows the relative performances of STASIS and LSA in calculating the STSS of the sentence pairs in the new dataset. In the table LSA results are re-scaled as discussed in chapter 2. Table 3.15 also re-scales the human ratings 0 – +4 to 0 – +1 and the ratings returned by STASIS have been rounded to 3 decimal places (the limit of accuracy of the individual human ratings). Correlation coefficients were calculated using the re-scaled and rounded forms.
Because both STASIS and LSA were sensitive to contractions (e.g. from do not to don't) all contractions were expanded before submitting the sentence pairs to the algorithms. Possessive forms (e.g. Tom's) were left unaltered on the ground that there is no simple expansion of the possessive.

Sentence Pair	Human Rating	STASIS Rating	LSA rating	Sentence Pair	Human Rating	STASIS Rating	LSA rating
S66	0.252	0.374	0.77	S98	0.056	0.266	0.57
S67	0.845	0.363	0.64	S100*	0.263	0.388	0.79
S68	0.782	0.2	0.75	S101	0.583	0.507	0.79
S69	0.024	0.366	0.63	S102	0.272	0.227	0.64
S70	0.71	0.428	0.63	S103	0.963	0.866	0.98
S71	0.939	0.444	0.94	S104	0.12	0.166	0.72
S72	0.553	0.531	0.64	S105	0.192	0.213	0.53
S73	0.455	0.365	0.71	S106	0.636	0.599	0.56
S74	0.181	0.395	0.7	S107	0.982	0.666	0.91
S75	0.363	0.354	0.53	S108	0.754	0.643	0.66
S76	0.784	0.609	0.89	S109	0.857	0.215	0.76
S77	0.032	0.177	0.62	S110	0.969	0.354	0.68
S78	0.546	0.261	0.56	S111	0.039	0.082	0.52
S79	0.978	0.332	0.97	S112	0.96	0.677	0.75
S80	0.125	0.104	0.58	S113	0.839	0.485	0.78
S81	0.895	0.328	0.87	S114	0.899	0.636	0.92
S82	0.607	0.717	0.66	S115	0.219	0.156	0.7

S83	0.952	0.658	0.95	S116	0.898	0.832	0.78
S84	0.512	0.391	0.81	S117	0.138	0.239	0.54
S85	0.428	0.328	0.82	S118	0.3	0.393	0.65
S86	0.023	0.188	0.5	S119	0.171	0.506	0.53
S87	0.769	0.775	0.94	S120	0.89	0.593	0.86
S88	0.766	0.669	0.71	S121	0.707	0.729	0.64
S89	0.805	0.392	0.78	S122	0.03	0.313	0.54
S90	0.186	0.383	0.64	S123	0.957	0.711	0.96
S91	0.021	0.263	0.73	S124	0.185	0.218	0.71
S92	0.558	0.574	0.89	S125	0.238	0.247	0.51
S93	0.426	0.657	0.75	S126	0.027	0.46	0.55
S94	0.491	0.187	0.61	S127	0.775	0.477	0.72
S95	0.312	0.491	0.55	S128	0.124	0.116	0.53
S96	0.42	0.182	0.63	S130*	0.04	0.103	0.5
S97	0.784	0.631	0.62	S131	0.764	0.405	0.82

Table 3.15 STSS ratings for the dataset from humans, STASIS and LSA

*The calibration sentence pairs, SP99 and SP129, should be omitted when using the new dataset to evaluate an algorithm. This is because the two sentence pairs will already have been used if the original dataset is used for preliminary testing (as SP64 and SP5 respectively).

The Pearson product-moment correlations obtained from these results (calculated in Minitab) were:

- LSA vs. Human ratings r = 0.693, P-Value = 0.000

- STASIS vs. Human ratings r = 0.636, P-Value = 0.000

Some interesting points arise from these results.

- The surprising higher performance of LSA has been shown to be consistent. This adds weight to the arguments of LSA's creators that it is modelling subtleties of the language.
- STASIS has not behaved as well as might be expected. This is discussed in sections 3.4.4 and 3.4.5.
- Both LSA and STASIS have lower correlations with the STSS-131 ratings than they did for STSS-65, although LSA is better than the worst human. This suggests that STSS-131 does indeed provide a more challenging representation of the language. One aspect of this is the representation of a wider range of dialogue acts.
- Because of its complexity and realism, STSS-131 is expected to present an enduring challenge capable of measuring improved performance of the new generation of STSS measures and those that follow.

3.3 The Discriminability Study

For the foreseeable future, there will always be the question of whether ground truth has been obtained in STSS experiments and new datasets will always be under scrutiny. Consequently it was decided to use an alternative technique from the field of Signal Detection Theory (Egan, 1975) for a subset of STSS-131 and examine the consistency of the ratings obtained. This technique, discriminability, measures semantic distance allowing the exploration of the questions:

- Are the ratings obtained from the participants consistent with each other across the two techniques for the whole sample being considered?
- If they are different, is there any evidence that one is superior to the other in approaching ground truth?
- If an individual pair of items is different, are there potential explanations which can shed light on ground truth similarity?
- How easy are each of the techniques for members of the general population to understand and apply?
- How efficient is the discriminability technique compared with card sorting in generating similarity ratings?

To fulfil the aim of using discriminability rating to measure STSS, three major phases were required:

1. conducting trials with participants to acquire a set of paragraph contexts for a representative sample from the 64 sentence pairs (Experiment 3.3)
2. design and conducting of a series of sentence-to-paragraph sorting trials to collect discriminability-based semantic distance ratings (Experiment 3.4)
3. conversion of semantic distance to semantic similarity and conducting a comparative evaluation of the results obtained from the two methods

SDT is concerned with detecting a signal in the presence of noise (Egan, 1975). Given a requirement to make a decision on whether or not a signal occurs in the presence of noise there are four possibilities: Hit (signal correctly classified as present), Miss (signal

incorrectly classified as absent), False Alarm (missing signal incorrectly classified as present) and Correct Rejection (missing signal correctly classified as absent).
The technique is applied in experiment 3.4 where the experimental task involves sorting which sentences belong with which paragraphs. A correct match is a Hit and an incorrect match is a False Alarm. The measure of discriminability, d', is calculated using Z-scores equation 3.2:

$$d' = z(H) - z(FA) \qquad (3.2)$$

where Z(H) is the Z-score of the hit rate and Z(FA) is the Z-score of the false alarm rate. The larger the value of d' the more discriminable the signal is from the noise.
An illustration of calculating a Z-score is given in appendix 3.14.
To conduct a discriminability study a number of experimental features and parameters must be determined. These are:

- a method of getting a participant to make the discrimination between the two sentences in a pair (determining the design of materials)
- determining the number of sentence pairs to be used in the study
- determining the number of participants to be used in the study

The method for measuring the discriminability of sentences in a pair was to create a set of "cloze paragraphs", each stimulated by one of the sentences in a pair. In a cloze paragraph, the stimulus sentence was replaced by a token. The discrimination task involved matching a cloze paragraph to its stimulus sentence when presented with both sentences in the pair. A correct match is a Hit and an incorrect match is a False Alarm, allowing d' to be calculated over the whole set of judgments. This method was specifically created for the work in this thesis and was derived from the "cloze sentence" technique used in word semantic similarity measurement (Miller and Charles, 1991, Charles, 2000).
The number of sentence pairs and the number of participants collectively determine the number of sortings (matching items to contexts) and hence the statistical power of a discriminability experiment. Increasing either increases the amount of participant effort required, meaning either more work on the part of individual participants or larger numbers of participants being required. Prior work on word semantic similarity provides some guidance for setting these parameters.

Miller & Charles (Miller and Charles, 1991) used 6 word pairs. The sorting task used 25 cloze sentences for each word and 24 participants. Consequently there were 600 context sortings per word resulting in 1200 for a pair. Each participant performed 300 sortings. The choice of 25 cloze sentences was justified on the grounds of the availability of suitable sentences in the Brown corpus.

(Charles, 2000) ran another series of SDT experiments, using the 30 word pair subset of the Rubenstein & Goodenough dataset. The sorting task used 15 cloze sentences for each word and groups of 9 and 24 participants sorting the contexts for 4 or 5 word pairs. Consequently there were 270 – 720 sortings per word pair.

Charles' experiment appears somewhat haphazard and suggests that using 30 word pairs resulted in an unmanageable size for the experiment. Choice of numbers of word pairs and participants for this work is described in the following sections.

3.3.1 Experiment 3.3 Creation of a set of paragraph contexts

Experiment 3.3 was a substantial task in its own right. It required the steps of:

1. selecting a sample of sentence pairs which are evenly distributed through the similarity range from the set of 64
2. designing a blocked study to distribute subsets of the materials (with equivalent cognitive load) to participants
3. specification of a suitable population sample
4. design of a data collection instrument (questionnaire) using individual stimulus sentences
5. running the sessions to collect the paragraphs

In experiment 3.3 a trial consisted of a participant writing a set of 32 paragraphs (2 for each sentence in a particular block). A session was an organised event at which a number of trials took place at the same time and place.

3.3.1.1 Materials - Selecting a sample of sentence pairs

The criteria for selecting the sentence pairs were keeping the sample to a manageable size (participant workloads) and a representation of the range of similarities. A secondary factor was to maintain the language feature representation criterion as far as possible. It was decided that 8 sentence pairs would give sufficient coverage by allowing for 2 low similarity, 2 medium-low similarity, 2 medium-high similarity and 2 high similarity pairs,

i.e. 2 pairs each from the ranges 0-0.99, 1.0-1.99,2.0-2.99 and 3.0-4.0 from the similarity scale used in the previous experiments.

Table 3.16 shows the sentence pair number, target (equal interval) similarity ratings, and those of the pairs actually selected. It also shows which properties from the complete dataset have been represented in this subsample.

SP	Properties	Similarity	Target Similarity N=64
126	Statement/Statement: two adverbs: low frequency : <degree partially> vs. <frequency mostly> also damp vs. burnt has elements of antonymy	0.11	0
124	Statement/Statement: two nouns abstract emotion: pleasure (+) vs. doubt (-)	0.74	0.57
95	Statement/Question: same class colour adjective: grey	1.25	1.14
93	Statement/Instruction: same concrete nonliving nonbiological noun: fire	1.7	1.71
72	Statement/Statement: same verb: State 47.8: continuous locative: cover: implied instruction	2.21	2.29
88	Statement/Statement: same manner adverb: safely: expected high similarity: paraphrased (alternative another ingesting verb)	3.06	2.86
120	Question/Question: A-P verb: (ingesting) 39.1 drink	3.56	3.43
107	Instruction/Instruction: same noun: concrete nonliving nonbiological: normally found outdoors: hill	3.93	4

Table 3.16 Selected sentence pairs (8)

The final column of the table, target similarities, represents the ideal similarity rating for each of the sentence pairs. Due to time constraints, the sentence pairs were selected on the basis of a substantial sample of returns but not the full N=64 combined sample. They are a little further from the desired values after including the whole sample, but the maximum deviation is 0.17, which is 4% of the full similarity range. SP88, at 3.06, was intended to be in the upper end of the medium-high similarity which had an upper limit of 2.99. This was accepted in view of the other properties that SP88 contributed to the set (shown in the properties column of table 3.16).

3.3.1.2 Creating a blocked design

The basic task involved in the data collection experiment was to write a paragraph containing a supplied sentence. A number of participants would be required to generate

sets of paragraphs covering all of the sentences in the experiment. These paragraphs would then have the original stimulus sentences deleted to create a set of cloze paragraphs for use in the discrimination task.

To inform the design, a review was made of the prior work on word similarity (Miller and Charles, 1991, Charles, 2000). However, there was some concern about making a priori deductions from this word-based work to a sentence-based approach – it was difficult to gauge the difference in effort require to produce cloze paragraphs compared to cloze sentences and to gauge the cognitive effort involved in sorting sentences to cloze paragraphs as opposed to sorting words to cloze sentences. The author of this thesis undertook a dry run of the processes involved to obtain some empirical basis for the necessary decisions on the workload.

As a consequence it was decided to use 32 participants in two blocks of 16 (designated blocks C and D). The 8 sentence pairs were split into 2 sets of 8 individual sentences; one set containing all of the first sentences in the pairs was allocated to block C and the other containing all of the second sentences in the pairs was allocated to block D.

This prevented the possibility of spurious semantic overlap (it would not be possible for two paragraphs written by the same person to be presented to participants in a discrimination task at a later stage in the study). Each participant was asked to write two paragraphs producing a pool of 32 paragraphs for each sentence, from which an appropriate sample would be selected.

Appendix 3.15 contains extracts from the paragraph data collection instrument.

3.3.1.3 Sampling the population

Having set a group size of 32, the nature of the population sample was considered. The task of generating the paragraphs was creative and potentially difficult to understand. Consequently a selective sample was specified to maximise the likelihood of useful paragraphs being returned. This sample consisted of a mix of students and non-student members of the general population who were strongly verbally oriented. The sample was strongly biased towards an Arts/Humanities background where verbal skills are most important. Most of the students were on degree programmes from these faculties, with A-levels in the appropriate area. Non-students largely had graduate level or better qualifications in a similar range of subjects and tended to work in educational areas with a bias towards English teaching, work in the theatre and the like.

It must be emphasised that a general population sample was not suitable for this task. Recruitment was selective, through recommendation by qualified individuals of potential participants with suitable language skills.

3.3.1.4 Design of the questionnaire data collection instrument

The basic task was to write a natural paragraph of between 3 and 5 sentences long containing a stimulus sentence (with instructions not to modify the stimulus sentence). An emphasis was placed on producing natural paragraphs as shown by the following extract from the instructions:

> "Above all, we are looking for natural paragraphs. Please write the kind of paragraph that you would be likely to use in a real conversation. If you can't do this try to think of a paragraph that you would naturally write in an Internet forum, an e-mail or a letter. Alternatively, try to think of a paragraph that someone might use with you in one of these circumstances. If all else fails try to think of a paragraph that someone would use on a radio or television programme that you would listen to."

Because the cognitive load involved in writing the paragraphs was unknown and difficult to predict a two stage questionnaire was produced. The first stage presented all of the 8 sentences allocated to the participant in a particular order, asking for the production of one paragraph for each sentence. The second stage invited participants who completed the first stage with time to spare, to make a second pass through the sentences in the same order, writing another different paragraph.

The two stages of the data collection questionnaire were identified as Section A and Section B within the questionnaire. The two groups of sentences were used to produce two different questionnaires, for blocks C and D. To prevent effects of ordering, two versions of each were produced. The first one had a randomised order of the 8 sentences (the same order being used in both sections A and B). The second version reversed the randomised order used in the first. Thus 4 questionnaires were produced for the data collection stage:

- Block C Forward
- Block C Reversed
- Block D Forward
- Block D Reversed

The general instructions were followed by a specific example:

"Please print (in BLOCK CAPITALS) a paragraph, between 3 and 5 sentences long in the boxes below, using the sentence:

He gave me a friendly warning against trying to cross by the old footbridge."

A grid was provided to write in the paragraph and the same format was followed for each of the sentences which appeared on separate, successive sheets.

3.3.1.5 Paragraph data collection trials

The non-students were uncompensated volunteers who completed the questionnaires unsupervised. Student volunteers were paid £5 compensation to attend a 1 hour session on University premises.

3.3.2 Experiment 3.4 Collecting discriminability ratings for the sentence pairs

Experiment 3.4, collecting the discriminablity-based semantic distance ratings, involved the following sequence of steps:

1. extraction of suitable and balanced sets of contextual paragraphs
2. design of a blocked study to distribute the materials to balance cognitive load amongst participants and prevent priming bias
3. design of the questionnaire instrument to collect judgements of sentence-to-context allocation
4. running the sessions to collect the judgements
5. calculation of discriminability ratings for the sentence pairs

In experiment 3.4 a trial consisted of a participant performing sentence-to-context allocations for 4 of the sentence pairs. A session was an organised event at which a number of trials took place at the same time and place.

3.3.2.1 Extraction of contextual paragraphs

It was decided to obtain 512 sortings for each sentence pair. This would involve 64 participants. To achieve this, 16 of the 32 paragraphs generated were extracted for each sentence, to be used in the sorting experiments. The dry run performed by the author was informative for the selection process and different criteria from Miller & Charles' random sampling and Charles' "representative" sampling were used.

Basically, all of the paragraphs were examined for potential "pitfalls" and those with pitfalls discarded. Random sampling was used where choice was available from surplus paragraphs which did not have pitfalls.

What is meant by a "pitfall?" One possibility is that a user could write a paragraph into which any sentence could fit meaningfully. Two examples found in the dataset were:

> Here's the sentence we've got to decode, it's crucial. <---------- *insert sentence here* ---------->Bring the solution with you. Wear this balaclava.

And

> What do you mean you don't understand? <---------- *insert sentence here* ---------- >It's as simple as that.

Their similarity arises in virtue of their both being English sentences and for no other reason. Whilst it is true that semantically similar English sentences have more in common with each other than say a sentence and a car, or a sentence and a slug, this similarity is meaningless in a study of semantic similarity – effectively it is low level noise (in a study with an extremely large sample of paragraphs it would have little impact apart from ensuring no sentence pair could be guaranteed to have an absolute zero similarity).

The second problem observed is that certain lexical features present in a sentence or paragraph may promote the use of a surface lexical feature approach to performing the sorting, rather than using the meaning of the sentences. One example was the occurrence of the word "meal" in the sentence:

> *Would you like to drink this wine with your meal?*

There were 6 occurrences of the word *meal* in the contextual paragraphs.

However, there were 9 occurrences of the word meal in the contexts for the paired sentence:

> *Will you drink a glass of wine while you eat?*

For each sentence in this pair, 4 paragraphs were selected that contained the word *meal* (one paragraph contained it twice) and participants were explicitly instructed not to use "crossword" or puzzle solving techniques instead of the actual meanings.

Appendix 3.16 contains a sample of the paragraphs produced by the block C participants.

3.3.2.2 Pilot study of the discriminability materials

This discriminability experiment was highly novel – although two experiments had been conducted on word discriminability, in the past there was no process for extrapolating from these to make inferences about sentence discriminability.

Also, the numbers of participants and sortings that were planned were believed to be adequate (based on the fact that 512 sortings fell within the range used by Charles (Charles, 2000) from 270-720 sortings). However, the fact remained that 512 was considerably less than the 1,200 sortings used in the earlier Miller & Charles study (Miller and Charles, 1991). Consequently it was felt that there was no margin for wastage or returned sortings due to process errors or blunders by participants, so a pilot study was conducted. The knowledge gained from the pilot study would minimise the chance of this happening.

The pilot study was conducted at a difficult time of the year as most students were away studying for exams. However, recruitment was performed using the local University chaplaincy system which acted as a social networking point for students.

In a normal experiment this could be criticised as a source of bias, as the students using the Chaplaincy are all Roman Catholics – however this pilot was solely concerned with collecting information about procedural matters such as clarity of instructions and timing, so it was not considered to be an issue. The data collected was not used in the statistical analysis.

In the pilot study, students were compensated at a rate of £5 to complete the questionnaire which was estimated to take 1 hour. They were allowed to complete in their own time with the usual instructions about completing it in an undisturbed single sitting. They were also asked to inform the investigator of the amount of time taken.

The materials were blocked in a variety of ways to investigate workload.

As well as the usual covering letter, extra information was also provided to assist unsupervised participants; this is shown in appendix 3.17.

The outcomes of the pilot were:

- the participants understood the task and found the instructions to be clear
- the participants were not unduly fatigued by the quantity of work
- different sentence pairs took different amounts of time process and this appeared to be correlated to the degree of semantic similarity (the more similar the greater the time taken on the sentence pair)
- the work took less than 1 hour even for the highest similarity sentence pairs

The conclusions were:

(a) confidence that the experiments could be carried out without wastage due to error

(b) a decision to produce two different combinations of materials for the sortings to even out the workload

(i) a combination of the highest and lowest similarity sentence pairs

(ii) a combination of the two medium similarity sentence pairs

(c) a decision to double the quantity of sortings performed by a participant in the final study compared with that in the pilot and compensate at £10 for up to 2 hours work.

Decision (c) was based on the fact that the pilot participants had not found the work to be too taxing. Doubling the task meant that each participant would process one sentence from pairs making up one half of the set (either (i) or (ii) in decision (b)).
It was felt that the benefit of this was that it would be easier to recruit student participants for a 2-hour session with a compensation of £10 as opposed to a 1-hour session with a compensation of £5. Finally, it was decided to inform participants that the experiment would require them to commit up to 2 hours of their time (in case the pilot participants were unusually speedy).

3.3.2.3 Blocked design of experiments to collect discriminability ratings

There were several reasons for using a blocked design, all of which were concerned with obtaining the best distribution of materials across the participants to prevent confounding effects.
In this experiment it was decided to use 64 participants, 32 students and 32 non-students. It is important to emphasise that although there is a serious element of between-groups design and pairing of students with non-students, there was no intention to use the results to explore subtle hypotheses about the differences between students and non-students. This would be interesting, but it would require double the number of participants (n=128) to produce results that would be of a reliability consistent with other results in this study. Evidence has been found from previous experiments in this study (section 3.2.4.5) to support the hypothesis that in tasks of this type, student and non-student samples can be

combined to provide a better picture of the general population and all of the blocking is aimed at getting the best quality data possible from these combined groups.
Blocking was intended to provide the following advantages:

- splitting the participants into 2 blocks of 32 resulted in two, manageable stages
- splitting each of the blocks of 32 into 16 students/16 non-students (pairing students and non-students in corresponding blocks of 32 ensured each combination of material was seen by 1 student and 1 non-student)
- "pair of sentence pairs" blocks of material were of approximately equal cognitive load
- combinations of blocks of materials could be arranged so no participant would be required to work with both sentences from the same pair

To even out the overall cognitive load, pairs of sentence pairs were specified according to their similarities as shown in table 3.16:

SP126 & SP107
SP124 & SP120
SP95 & SP88
SP93 & SP72

For example, SP126 and SP107 is a combination of the lowest and highest similarity sentence pairs in the table. These pairs were then combined as:

SP126-107 & SP 124-120
And
SP95-88 & SP93-72

The blocked design is shown in appendix 3.18.

3.3.2.4 Design of questionnaire instrument to collect sortings for discriminability measurement

There were two sets of material for each block 32 participants, a combination of Sentence Pairs 126, 107, 124 and 120 OR a combination of Sentence Pairs 95, 88, 93 and 72.

Each individual sentence pair set came in two versions, a randomised order (F = forward) and the reverse of the original randomised order (R = reversed). For convenience of organising materials a combination of materials for two sentence pairs would have them same version within each sentence pair, so SP124-120R, for example, refers to a combination of materials for 8 pairs from SP124 in their reversed order, followed by a combination of materials for 8 sentence pairs from SP120. Four combinations were created for each package of two sentence pairs e.g.

SP120-124F
SP124-120F
SP124-120R
SP120-124R

These packages were paired in various combinations of forward and reverse ordering of sentences and order of presentation of the two packages. A full matrix of blocking combinations used is shown in appendix 3.18.

In summary, each sentence pair contributed 16 sortings to the overall work, eight involving cloze paragraphs derived from one sentence in the pair and 8 involving paragraphs derived from the other sentence.

The two blocks of 32 participants used two mutually exclusive sets of 8 paragraphs taken from the 16 paragraphs for each sentence extracted from the set generated in section 3.3.1.

Finally, the student and non-student participants were interleaved through the two blocks of 32 and the second block was staggered with respect to the first, the outcome being that for each particular combination of materials, e.g. SP120-124F &
SP107-126F, the combination I the first block would be allocated to a student and that in the second block to a non-student, and vice versa.

So, allowing for the fact that each of the two 32 blocks used different sets of paragraphs derived from the sentences, each participant received a unique permutation of the materials. This was facilitated by the approach taken to generating the work component packages.

Having made up each complete set of materials, instructions were added at the start and participant details were requested at the end. Because of the length of process (and again facilitated by the work packaging approach) a summarised repetition of the instructions was inserted halfway through the questionnaire to refresh the participants' memories.

3.3.2.5 Specification, selection and recruitment of a suitable population sample

The basic intention was to recruit a group of participants broadly comparable to those used in the card sort similarity rating experiment (experiment 3.2) described in section 3.2.4. The time at which the experiments were conducted fell partly during the student vacation period. Consequently, the student sample included some who were recruited whilst at home on vacation, so a slightly broader range of University backgrounds was covered. Also a number of non-student participants were parents of the students who volunteered. The demographic principles were the same as in experiment 3.2:

- to recruit student participants in the 18-22 age band, and non-students distributed over the range from 21 upwards
- to obtain a balanced representation of both genders in both students and non-students
- to obtain a balance of Science/Engineering vs. Arts/Humanities backgrounds in both groups

3.3.2.6 Experimental trials

The covering letter and ethics statement for the experiment were taken from the pilot study. The instructions are in appendix 3.19 and an extract from the discriminability rating form is in appendix 3.20.

The majority of the undergraduates completed the task at one of a number of supervised sessions organised in their faculties. A small proportion who completed the task during a vacation period did so in the same manner as the non-student volunteers. Non-student volunteers from the general population completed in their own time following the usual instructions to set aside an hour when they would be undisturbed to focus exclusively on the work.

The pilot discriminability study provided evidence that the discriminability task was sufficiently clear to be performed by members of the general population without supervision and no problems were reported by the participants. Again, running group sessions at university premises for the majority of students was convenient from the administrative point of view.

3.3.2.7 Results of discriminability experiment

As shown in table 3.17, the age criterion was achieved with an acceptable possibility of overlap around the age of 22.

A good educational balance between Arts/Humanities and Science/Engineering was obtained in the students group and there was a substantial proportion of people in the non-students with a non-specialised level of education, as shown in table 3.18.
Overall, the objectives of obtaining a comparable sample with the card sort was achieved, with the proviso that this sample had fewer people with university education and slightly more with a further education qualification.

Students		Non-students	
Age band	Number	Age band	Number
< 18	0	<21	0
18-22	23	21-30	5
>22	3	31-40	9
Withheld	6	41-50	6
		51-60	7
		>60	5
		Withheld	0

Table 3.17 Age distributions for Student and Non-Student population samples

Students		Non-students	
Degree discipline	Number	Highest qualification	Number
Arts/Humanities	13	None	5
Science/Engineering	18	GCE/GCSE	5
Interdisciplinary	1	Further Education	4
Withheld	0	BSc	4
		BA	6
		MA	1
		Professional Qualification	3
		MD (Doctor)	1
		Withheld	3

Table 3.18 Educational background distributions for Student and Non-student samples

As can be seen from table 3.19, a good balance of representation of genders both within groups and between groups was obtained.

Gender	Number in Students	Number in non-students
M	15	12
F	16	14
Withheld	1	6

Table 3.19 Gender distributions for Student and Non-Student population samples

The results of the experimental runs are tabulated in the following contingency tables (3.20 – 3.27). These results are used to calculate the probability of a hit and the probability of a false alarm for one sentence from the pair in the presence of the other, over the set of paragraphs for a particular sentence pair and over all of the participants who sorted that sentence pair (256 sortings for paragraphs allocated to each sentence in the pair, giving 512 sortings in total).

Z-scores were then obtained for these two probabilities by subtracting the z-score for the probability of a false alarm from the z-score of the probability of a hit as described in appendix 3.14.

Sentence Pair 72

A: You shouldn't be covering what you really feel.

B: There is no point in covering up what you said, we all know.

Stimulus	Response A Class 1	B Class2	Sum
Class1 A	217	39	256
Class2 B	68	188	256
Sum	285	227	512

Table 3.20 Contingency Table for SP72

So to interpret table 3.20, cloze paragraphs that originally contained sentence A were allocated to it 217 times and incorrectly allocated to sentence B 39 times, correctly adding up to 256 paragraphs originally derived from sentence A. The class column total show a response bias towards sentence A of 285 allocations vs. 227 for sentence B and that this arose because there was a greater number of misclassifications of sentence B as sentence A than vice versa.

Sentence Pair 88

A: We got home safely in the end, although it was a long journey.

B: Though it took many hours travel, we finally reached our house safely.

Stimulus	Response A Class 1	B Class2	Sum
Class1 A	157	99	256
Class2 B	107	149	256
Sum	264	248	512

Table 3.21 Contingency Table for SP88

Sentence Pair 93

A: I need to dash into the kitchen because I think my chip pan is on fire.

B: In the event of a chip pan fire follow the instructions on the safety note.

Stimulus	Response A Class 1	B Class2	Sum
Class1 A	246	10	256
Class2 B	1	255	256
Sum	247	265	512

Table 3.22 Contingency Table for SP93

Sentence Pair 95

A: I'm going to buy a grey jumper today, in half an hour.

B: That's a nice grey top, where did you get it from?

Stimulus	Response A Class 1	B Class2	Sum
Class1 A	235	21	256
Class2 B	29	227	256
Sum	264	248	512

Table 3.23 Contingency Table for SP95

Sentence Pair 107

A: Meet me on the hill behind the church in half an hour.

B: Join me on the hill at the back of the church in thirty minutes time.

Stimulus	Response A Class 1	B Class2	Sum
Class1 A	160.5	95.5	256
Class2 B	155	101	256
Sum	315.5	196.5	512

Table 3.24 Contingency Table for SP107

Sentence Pair 120

A: Would you like to drink this wine with your meal?

B: Will you drink a glass of wine while you eat?

Stimulus	Response A Class 1	B Class2	Sum
Class1 A	181	75	256
Class2 B	77	179	256
Sum	258	254	512

Table 3.25 Contingency Table for SP120

Sentence Pair 124

A: The pleasure that I get from studying, is that I learn new things.

B: I have a doubt about this exam, we never got to study for it.

Stimulus	Response A Class 1	B Class2	Sum
Class1 A	250	6	256
Class2 B	7	249	256
Sum	257	255	512

Table 3.26 Contingency Table for SP124

Sentence Pair 126

A: The damp was mostly in the very corner of the room.

B: The young lady was somewhat partially burnt from the sun.

Stimulus	Response A Class 1	B Class2	Sum
Class1 A	254	2	256
Class2 B	0	256	256
Sum	254	258	512

Table 3.27 Contingency table for SP126

The data from tables 3.20 – 3.27 was used to calculate d' values. These values actually correspond to semantic distance between the two sentences comprising a pair. Under many circumstances it is feasible to plot d' directly against a measure of similarity and look for an inverse correlation. However, d' asymptotically approaches infinity as z(H) – z(FA) approaches 0 and the z-scores for the false alarms becomes too large to calculate in the case of SP126. As an experiment, replacing the 0 value for false alarms for sentence B with a small number (and progressively smaller number still) showed a trend of increasing size of d'.

Consequently, a well-known function (equation 3.3) for transform ranges from $0 \rightarrow \infty$ to $1 \rightarrow 0$ was applied which transformed the d' measure into a measure of similarity and solved the problem of asymptotic behaviour:

$$d'\,\mathrm{Sim} = \frac{1}{1 + d'} \qquad (3.3)$$

The d' similarity measures are shown, along with the card sort measures and substituted ranks in table 3.28. The relationship between the d' similarity measures derived from SDT and the similarity measures obtained from card sorting are shown in the graph, figure 3.15. Visually, the two measures appear to correlate reasonably well, with two points bucking the general trend – SP95 and SP120. Statistical measures show a good correlation between the two similarity measures.

Sentence Pair	d' Similarity	Card Similarity	d' Similarity Rank	Card Similarity Rank
126	0	0.11	7	8
124	0	0.74	7	7
95	0.277716	1.25	5	6
93	0.184443	1.7	5	5
72	0.376998	2.21	4	4
88	0.668908	3.06	2	3
107	0.946731	3.93	1	1
120	0.483819	3.56	1	1

Table 3.28 d' vs. Card Sort similarities (with ranks)

The Pearson product-moment for the two measures is 0.924 ($p = 0.001$) and the Spearman rank correlation coefficient is 0.946 (calculated using Minitab). These are high correlations indicating very good agreement between the similarities obtained from the two techniques. It could be argued that the sample is too small for the result to be significant. However support for the value of the results comes from three sources. Firstly, the p-value for the Pearson product-moment correlation coefficient is well below the generally accepted limits of 0.05 and 0.01 for significance of the correlation. Secondly, each of the 8 data points does not represent a single measurement. The card sort data is derived from 32 ratings each from an individual participant and the d' is derived from 512 sortings performed by 32 participants. Finally, the Spearman rank correlation coefficient, which is considered to be more robust to experimental variations, actually gives the higher correlation.
Minitab regression analysis comes up with a best fit equation for the two measures of:

$$d' = -0.0911 + 0.221*\text{Card} \qquad (3.4)$$

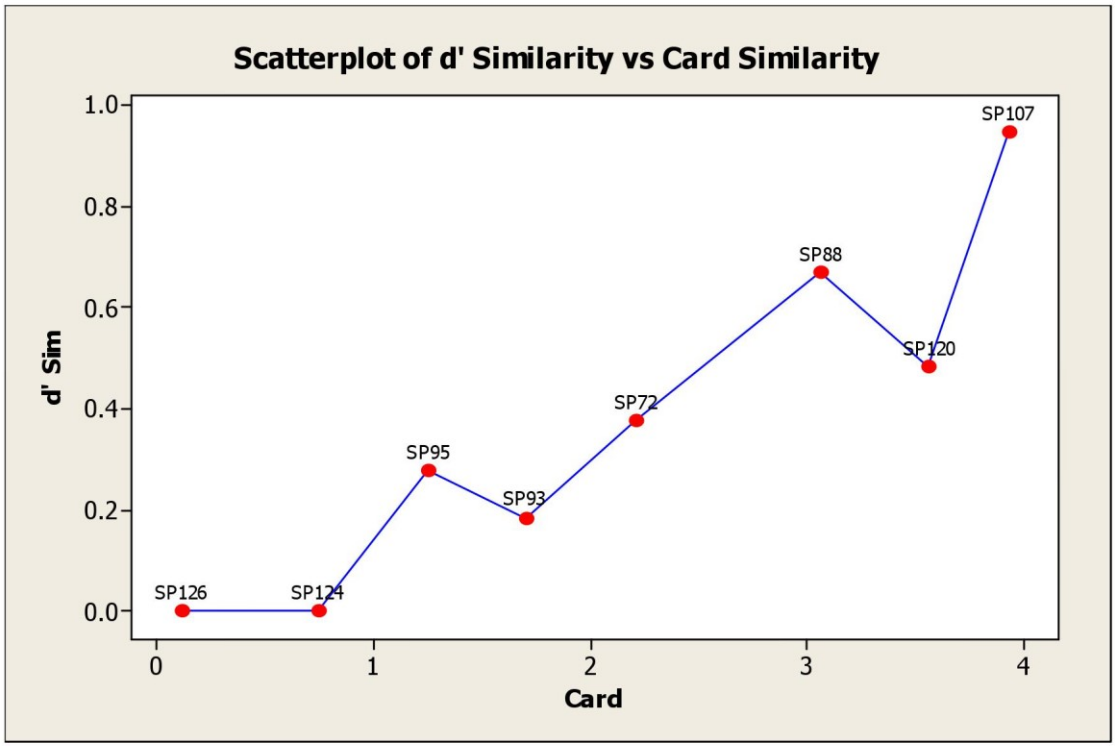


Figure 3.15 Scatterplot with line graph of d' similarity vs. Card Sort similarity

It should be noted that correlation is the key issue as linear transformations of the measurements are permissible (as in transforming temperatures Fahrenheit to Celsius).

Two cases did not fit the trend and these deserve further consideration. These were SP95 and SP120.

SP95

I'm going to buy a grey jumper today, in half an hour.
That's a nice grey top, where did you get it from?

This sentence pair mixes two dialogue acts, an assertion and a question. Considering the other sentence pairs, SP72 combines an instruction with an assertion, but the assertion can also be interpreted as an indirect instruction Dialogue Act. SP88, SP124, and SP 126 are all pairs of assertions. SP93 is an instruction paired with an assertion. SP120 is a pair of questions and SP107 is a pair of instructions.
The possible explanation is that the two techniques are influenced in different proportions by dialogue act vs. propositional similarity. It is not possible to determine which value is

the better representation of the ground truth similarity between the two sentences in the pair, but does suggest the factors influencing STSS judgement should be studied further.

SP120

Would you like to drink this wine with your meal?
Will you drink a glass of wine while you eat?

Prima facie, the card sort rating of 3.56 (0.89 when scaled from 0 – 1) is more plausible than the SDT similarity of 0.483819. The possible explanation in this case is that there were surface lexical features in the paragraphs that the participants were able to use in place of the meanings of the sentences to perform the sorting activity. This particular pair was the original reason for the scrutiny of the selected paragraphs to avoid unusual features and balance others that might give clues or mislead. However, the strategy may have been unsuccessful in this case.

The conclusions from this comparison between the two techniques are that each technique provides independent evidence to support the validity of the other due to the high Spearman rank correlation coefficient ($\rho = 0.946$). Also, some variation in measured similarities may be due to the different influence of the factors semantic similarity and propositional content. Finally where there is variation it is not obvious which method is getting closer to ground truth and SDT is not the "Magic Bullet" for capturing human intuitions about semantic similarity that it first appeared to be.

3.4 Discussion

This chapter has described the creation of a new benchmark dataset for STSS algorithms, STSS-131 which reaches the highest possible quality obtainable at the current state of the art.

3.4.1 Distinguishing features of STSS-131

Unlike any preceding work, every item in the dataset has a specific reason for being there in representing properties of the English language. Although it is not possible to cover the language exhaustively in such a small dataset (64 sentence pairs), the description of the work conducted in section 3.2 provides grounds to support the assertion that no better

coverage of the language could have been obtained in a dataset this size given present knowledge across the various contributing fields.

In the key areas of obtaining the human ratings in this chapter, care has been taken to obtain balanced and representative population samples far in excess of that taken by other researchers whose work has become prominent in the related field of word semantic similarity.

In generating materials, participants were been chosen on the basis of aptitude to produce English language utterances and paragraphs that would be realistic as components of dialogue. This care was rewarded and examination of the appendices for this chapter will reveal challenging but realistic data items with none of the unrealistic quirkiness of items such as "Time flies like an arrow; fruit flies like a banana" (Groucho Marx) which are so popular in the Computational Linguistics literature (Palmer, 2006).

It is expected that STSS-131 will be adopted for use by the research community on a wider basis to evaluate the emerging STSS algorithms in the same way as STSS-65.

3.4.2 The discriminability study

The second major component of this chapter involved conducting an independent study using the Signal Detection Theory technique of discriminability. Although two studies have been conducted using SDT for word similarity (Miller and Charles, 1991, Charles, 2000), it has not been adopted on a larger scale in semantic similarity studies because of its high labour intensiveness – which means large numbers of participants have to be recruited and careful blocked designs used to obtain sufficient ratings of good quality. Again adapting this technique from word to short-text similarity was totally novel and the feared "gearing up effect" on the labour intensiveness of the technique in creating paragraphs occurred. However it did turn out that the rating process was less severely affected. It was only feasible to obtain ratings for 8 sentence pairs distributed approximately evenly across the similarity range (as measured using early returns from the card sort experiment), nevertheless this is an improvement on the 6 word pairs used in a corresponding word semantic similarity experiment (Miller and Charles, 1991).

3.4.3 Motivation for discriminability study

The discriminability study was conducted for two reasons. First, for the same reason as its use in the word similarity studies, to see if it corroborated or refuted the card sort study, that formed the basis of this chapter. The second was to investigate its feasibility and effectiveness for use on a larger scale. During the literature search it was intriguing that

this technique seemed to be the magic bullet – a way of obtaining similarity ratings through a simple task that would reduce subjectivity and eliminate the hazard of introspection on the part of participants.

In regard to the first reason, it did give good corroborative evidence that the similarity measurements being obtained. At 0.924, the Pearson product-moment correlation between the discriminability ratings and the card sort ratings falls about halfway between the performance of the average human ($r = 0.891$) and the best human ($r = 0.951$) in agreeing with the rest of the group.

In regard to the second, although the technique proved easy for participants to understand, opening up the way for a broader sample of the population to participate in experiments without supervision, some problems were found with contextual paragraphs as described in section 3.3.2.1.

The anomalously high discriminability (corresponding to low semantic similarity) of SP120 (*Would you like to drink this wine with your meal?* vs. *Will you drink a glass of wine while you eat?*) suggested that some other features (possibly lexical surface features) were being exploited by participants in place of sentence meaning in making some of their judgements. i.e. although they were asked to make their judgements using semantic content, their desire to make the correct assignment led them to use non-semantic features. Consequently the conclusion was that both discriminability and card sort techniques were generally valid and that the different insights they brought to semantic similarity measurement did not result in either being decisive. Discriminability measurement provided useful evidence to support the validity of the card sort measurement but was not a magic bullet that should replace it.

Further evidence of the validity of the process came from the two calibration sentence pairs. The ratings in the two different data sets were found to be not significantly different using Student's t-test, providing evidence for robustness of human similarity perception in the presence of contextual variation.

The final piece of evidence concerning validity comes from the Pearson product-moment correlation of 0.924 between the two population samples (students vs. non-students) provided evidence for robustness of perception whilst varying a combination of factors such as age, cultural background etc.

3.4.4 Comparison of STASIS and LSA using STSS-131

In the illustrative application of the benchmark dataset to comparing STASIS and LSA, the new dataset has met the objective of providing a more demanding test. The performance

of LSA supports the assertion that its underlying language model is more sophisticated than a mere bag of words. Somehow, LSA seems to be inferring at lease some of the structural information held in the function words from its content-word only vectors. Another factor may be that a number of concepts in the sentences are conveyed by phrases rather than single words. Where a phrase in one sentence has high semantic similarity to a word in another this will be missed by STASIS. LSA, on the other hand, would be expected to exploit this.

It must also be borne in mind that the parameter δ in STASIS was originally set empirically, on the basis of limited evidence (described in chapter 1). Therefore one possibility would be to investigate optimising δ to see if better settings could be found (combining STSS-65 and STSS-131 would give sufficient data for optimisation with a genetic algorithm using n-fold cross validation).

Other approaches, using alternative combinations of the word order and semantic similarities were also investigated briefly to see if STASIS could be improved. First a regression test was conducted in Minitab which produced the linear combination in equation 3.5:

$$\text{Overall Similarity} = 0.0938 - 0.014*\text{Order} + 0.974*\text{Semantic} \qquad (3.5)$$

When the overall similarities calculated using equation 3.5 were correlated with the human ratings, a correlation coefficient of 0.670 was obtained (P-Value = 0.000), which was an improvement, but still markedly below the LSA value.

Some earlier experiments in combining STASIS and LSA suggested that non-linear approaches may be more appropriate when combining different models of similarity[1], so a second experiment was performed using an Artificial Neural Network created in Matlab to combine word order and semantic similarities in place of the δ parameter in STASIS. The normalised results were re-scaled to bipolar form and a network trained using the Tansig transfer function. A network with one hidden layer containing 10 units combined the order and semantic similarities in a method which produced a set of similarity ratings with a Pearson product-moment correlation of 0.731 (P-Value = 0.000), which was markedly higher than LSA.

---

[1] This work was presented at the KES-AMSTA 2008 conference in an extended discussion of the paper "A Comparative Study of Two Short Text Semantic Similarity Measures" O'SHEA, J. D., BANDAR, Z., CROCKETT, K. & MCLEAN, D. (2008) A Comparative Study of Two Short Text Semantic Similarity Measures. *Lecture Notes in Artificial Intelligence*.

3.4.5 Weaknesses of current STSS measures

Investigating these problems with STASIS provided the stimulation for the next chapter. Basically, STASIS combines two numerical measures: Semantic and Word Order. The Semantic component measures similarity based on what the sentences are **about**. The Word Order component is supposed to add information about what the two sentences **say** about the semantic content.

However, it is possible that simply knowing about the positions of matching function words in two sentences is not enough to exploit the structural information conveyed by them. It is possible that a higher-level classification of sentences, by the Dialogue Act, could provide the necessary information to modify the Semantic measure.

However, this will be a moot point unless the Dialogue Act of a sentence can be classified accurately and rapidly (classification must not be computationally intensive if the resulting code is to be deployed on web-based Conversational Agents).

Therefore chapter 4 involves the creation and validation of a method of classifying Dialogue Acts using only function word features as input to AI classifiers.

Chapter 4 Towards a New Model of Short Text Semantic Similarity

4.1. Introduction

This chapter describes two problems faced by current STSS measures. These problems were underlined by experience from the creation of the benchmark datasets.

- The first is the fact that they do not take account of the Dialogue Acts of the two sentences being compared (discussed in chapters 2 and 3).
- The second is that they do not take account of the valence of the two sentences being compared (this was discussed briefly in terms of positive and negative affect in chapter 3).

After a more detailed discussion of the relevant issues, this chapter makes the following contributions:

- a proposed new model of Short Text Semantic Similarity
- a new method of Dialogue Act classification
- an experimental evaluation of the new method including
    - a new method of encoding features for Dialogue Act recognition
    - a series of experiments using Questions and Non-Questions to test the method
    - production of datasets for the experiments which will be made available to the research community

4.2 The Dialogue Acts problem

4.2.1 Definition of a Dialogue Act

The concept of DAs originates in the work of J.L Austin (Austin, 1975) which was further developed by John R. Searle (Searle, 1969, Searle, 1999). Although the term "Dialogue Act" is derived from Speech Act theory, it is specifically associated with machine-based dialogue management.

The DA separates out the point or purpose of the act from the propositional content. Fundamentally, "propositional content" means what the utterance is about, so the statement "the door is open" is about the door being open.

As an example of DAs (Green, 2007), starting with the propositional content the *door is shut*, DAs such as the question "Is the door shut?", the instruction "Shut the door!" and the assertion "The door is shut." can be derived.

Searle (Searle, 1999) extended work on DAs and developed the following classification (a few examples have been included as bullet points):

1. Assertive (to commit the hearer to the truth of the proposition)
    - statements
    - descriptions
2. Directive (to try to get the hearer to behave in such a way as to make his behaviour match the propositional content of the directive)
    - orders
    - requests
3. Commissive (A commitment by the speaker to undertake the course of action represented in the propositional content)
    - promises
4. Expressive (to express the sincerity of the speech act)
    - apologies
5. Declarations (to bring about a change in the world by declaring it to have been changed)
    - pronounce marriage
    - declare war

One complication is the Indirect Dialogue Act, in which one DA is performed indirectly by performing another e.g. "I promise to come and see you" combines Declarational, Commissive and Assertive.

Searle observed that there are a vast number of different DAs (Searle, 1999), so it is understandable that there are a number of different taxonomies breaking down DAs into different levels of granularity as well as covering slightly varied subsets of acts.

4.2.2 Dialogue Acts and STSS measurement

Existing STSS measures are effectively incapable of distinguishing between DAs because they focus on Propositional Content. This is illustrated for LSA (Deerwester et al., 1990) and STASIS (Li et al., 2006) in tables 4.1 and 4.2 respectively. The same propositional content is used in 3 sentences (examples derived from (Green, 2007)):

- The statement - "The door is open."
- The question - "Is the door open?"
- The instruction - "Open the door."

Short Text Pair	DA	LSA Rating
The door is shut.	Statement	1.0
Is the door shut?	Question	
The door is shut.	Statement	1.0
Shut the door!	Instruction	
Shut the door!	Instruction	1.0
Is the door shut?	Question	

Table 4.1 LSA similarity ratings for various combinations of Dialogue Acts

Short Text Pair	DA	Order Similarity Rating	Word Similarity Rating	STASIS Overall Rating
The door is shut.	Statement	0.77058	1.0	0.96559
Is the door shut?	Question			
The door is shut.	Statement	0.45767	0.98452	0.90549
Shut the door!	Instruction			
Shut the door!	Instruction	0.64194	0.98452	0.93313
Is the door shut?	Question			

Table 4.2 STASIS similarity ratings for various combinations of Dialogue Acts

Even though the sentences have completely different meanings because of their DAs, they are rated as identical by LSA and virtually identical by STASIS.

This is not a surprise if the descriptions of the two algorithms from Chapter 1 are recalled. LSA throws away all of the words in its stoplist (largely function words) so that the only information left in each of the sentences is the pair of words *shut* and *door*. STASIS, on the

other hand, uses all of the words and is sensitive to word order. So the first pair which consists of identical sets of words (*the door is shut.* And *is the door shut?*) has an overall similarity of less than one. But because word order is given a low weighting, it is still very close at 0.96559. Altering the parameter δ could help. However, a more realistic solution is to measure the similarity between the propositional content of the two sentences and then modify it based on the particular combination of DAs. This requires a method of identifying the DA of a sentence and some algorithm or parameters to make the modification.

4.3 The Valence Problem

4.3.1 Definition of Valence

"Valence" is a term originally used by chemists (Little et al., 1983) which has been adopted by researchers in emotion in the fields of cognition (Ortony et al., 1990), psychophysiology (Inaba et al., 2005) and agent-based systems (Oliveira and Sarmento, 2002).

Valence concerns positive or negative judgements of events, agents and objects giving rise to emotions. For example, Inaba used Valence to describe the positive or negative qualities of particular words in an ERP-base word recognition experiment (Inaba et al., 2005).

Once the concept of valence has been recognised, it can be seen that there are two variants. The first is the polar opposite (or bipolar) view; the second is the Boolean (or unipolar) view.

The bipolar view is illustrated by the pair T*he door is always shut* / T*he door is never shut*. There are degrees of openness between the extremes such as *the door is kept shut at weekends*; *the door is half open* etc. A suitable scale for recording the similarity of sentence pairs of this type would cover the range from *identical in meaning* (+1) through *completely different in meaning* (0) to *opposite in meaning* (-1).

In the unipolar variant the propositional content can be either true or false, for example *The door is shut* / *The door is not shut*. In this Boolean view any degree of openness counts as not shut.

4.3.2 Valence and STSS measurement

The problem is that existing STSS measures are relatively insensitive to polarity, in tables 4.3 and 4.4 both LSA and STASIS report high similarity for the opposite ends of the bipolar scale:

Short Text Pair	DA	LSA Rating
The door is always shut.	Positive extreme	0.95
The door is never shut.	Negative extreme	

Table 4.3 LSA result for opposites

Short Text Pair	DA	Order Similarity	Word Similarity	STASIS Overall
The door is always shut.	Positive extreme	0.58743	0.82302	0.78768
The door is never shut.	Negative extreme			

Table 4.4 STASIS result for opposites

In this case, the similarity ratings produced by the two algorithms are reasonable in terms of the modulus, but has the incorrect sign. Changing the sign from positive to negative would produce reasonable similarity ratings on the proposed -1 to +1 scale.

The unipolar case is exemplified by a statement and its contradiction. Again an example of the treatment of a unipolar pair by existing measures, using examples derived from (Green, 2007), is instructive.

Short Text Pair	DA	LSA Rating
The door is shut.	Statement	1.0
The door is not shut.	Contradiction	

Table 4.5 LSA ratings for a statement and its contradiction

Short Text Pair	DA	Order Similarity	Word Similarity	STASIS Overall
The door is shut.	Statement	0.66667	0.97238	0.92653
The door is not shut.	Contradiction			

Table 4.6 STASIS ratings for a statement and its contradiction

Table 4.5 illustrates that the LSA algorithm is completely unable to distinguish between the statement and its contradiction. Table 4.6 shows that STASIS fares a little better, as it detects some difference largely due to word order, but like LSA it does not capture the true intent of the speaker.

Processing sentences of this type can be important in real-world CA applications which may involve manipulative, face-saving or adversarial behaviour on the part of the human user (Crockett et al., 2009). An example of a face-saving hedge might be:

Agent: Did you pay on time?

Human: Well, I wouldn't say I paid late.

In a similar manner to DAs a solution would be to measure the similarity between the propositional content of the two sentences. Then classify the polarity relationship (unipolar or bipolar) and adjust the propositional content accordingly.

Bipolar relationships would be indicated by adjectives or adverbs (as discussed in chapter 3) and unipolar ones by function words such as *not, always, never* etc.

4.4 A Factor-based model of Short Text Semantic Similarity

This chapter proposes a solution to the problems of DAs and Valence in the form of a factor-based model. Factor-based models are well established in many fields; one of the best known is the 5-factor model of personality used in psychology (Norman, 1963).

4.4.1 The proposed model

The proposed new factor-based STSS model is based on the hypothesis that short text semantic similarity can be separated into a set of clearly defined factors which can be measured independently and then synthesised into an overall measure.

The general concept of the model can be described by equation 4.1:

$$Sim = \prod_{i=1}^{n} f(S_i) \qquad (4.1)$$

where S_i is a particular similarity factor.

Use of a multiplicative model allows each factor to play a full role in producing an output which is scaled in a true cosine range of -1 to +1. This range represents identical meaning

between a pair of sentences as +1, complete oppositeness as -1 and unrelatedness of meaning as 0.

In this work the model has three factors which are S_{Prop} (propositional content similarity), S_{DA} (similarity between DAs) and S_{Val} (similarity between valences). Accordingly, the equation becomes:

$$Sim = f_P(S_{Prop}) \times f_d(S_{DA}) \times f_v(S_{Val}) \qquad (4.2)$$

Where $0 <= S_{Prop} <= +1$, $0 <= S_{DA} <= +1$ and $-1 <= S_{Val} <= +1$. The weighting functions f_p, f_d and f_v may be used to tune the contributions of the different factors.

4.4.2 Requirements for developing the proposed model

Implementing the complete model requires the development of 3 crucial components, these are:

1. a component for measuring proposition content similarity
2. a component for calculating dialogue act similarity
3. a component for calculating valence similarity

It will also require appropriate parameters or functions to combine the factors. Implementation of the full model is beyond the scope of this thesis. However, existing STSS measures basically measure propositional content and adapting one of them for this role should be straightforward. Devising parameters to combine the factors is not practically useful until all 3 components have been developed. Therefore there was a choice between developing components for either Dialogue Acts or Valence as the next stage of the work in this thesis. Dialogue Act classification was chosen as the task for this chapter.

4.5 Development of the Slim Function Word DA classifier

DA classification is essential for measuring DA similarity, furthermore classifying DAs makes the most fundamental aspect of measuring DA similarity feasible immediately, because a pair of DAs can be classified as identical or different.

It was decided to investigate the feasibility of the factor-based model by developing one of the classifiers that would be required. The literature search in chapter 1 illustrated the

importance of the field of question answering across a whole range of IR activities. The established approach to question answering (Hirschman and Gaizauskas, 2001) absolutely takes it for granted that some previous process has identified an utterance as a question. Further searching established that recognizing a question is part of a much smaller field, DA classification (discussed below). Also, little consideration has been given to the computational overheads of DA recognition in a scalable web-based CA. The following sections introduce a new, efficient approach to DA classification, the Slim Function Word Classifier (SFWC). Designing this classifier required some consideration of existing approaches to DA classification.

4.5.1 The Classification of Dialogue Acts

There are three important components of DA classification:

- DA Taxonomies
- DA Feature Extraction
- DA Classifiers

4.5.1.1 Dialogue Act Taxonomies

No single, definitive DA taxonomy has emerged from prior research. However, there is a comprehensive model called DAMSL – Dialogue Act Markup In Several Layers from the University of Rochester (Allen and Core, 1997). DAMSL is a methodology, described as "a system for annotating dialogues" and the DAMSL manual comments "For any particular project, we would expect that the annotation scheme would be refined to provide further detail on phenomena of interest." (Allen and Core, 1997). It has clearly specified upper layers; the flexibility of the lower layers is intended to allow it to be adapted effectively in diverse domains, whilst allowing some prospect of interoperability or re-usability between projects. The highest level contains 4 categories of tags. These represent independent dimensions, so an utterance could be tagged with properties from all four categories. These dimensions are: *Communicative-Status, Information-Level, Forward-Looking Function* and *Backward-Looking Function.*

- *Communicative-Status* is only applied to uninterpretable utterances and includes the tags *Uninterpretable, Abandoned* and *Self-talk.*
- *Information-Level* tagging is described providing an abstract characterization of the content of the utterance. It includes the tags *Task, Task-management, Communication-management* and *Other-level.*

- *Forward-Looking Function* tags are concerned with the effect an utterance has on the subsequent dialogue and interaction. Categories include *Statement, Influencing-addressee-future-action, Info-request, Committing-speaker-future-action, Conventional Opening Closing, Explicit-performative, Exclamation,* and *Other-forward-function.*
- *Backward-Looking Function* tags are concerned with how the current utterance relates to the previous discourse. Categories include *Agreement, Understanding, Answer* and *Information-relation.*

In the case of *Forward-Looking Function* and *Backward-Looking Function* the categories can contain a number of different kinds of tags. *Forward-Looking Functions* are perhaps the most familiar with the *Statement* category containing tags for *Assert, Reassert* and *Other-statement.* In the *Backward-Looking Functions*, *Understanding* contains a subcategory *Signal-understanding* which itself contains 3 tags – *Acknowledge, Repeat-rephrase* and *completion*.

Following the DAMSL manual to the lowest level of detail, it is possible to identify 30 tags; however some of these are sufficiently unspecified to allow further decomposition, for example the *Information-Level, Other-level* category is described as capable of containing jokes, non-sequiturs and smalltalk.

The consequence of all this is that it is possible to find studies influenced by the DAMSL philosophy which either cut down the number of tags (and therefore DA categories) or expand them. Some examples are discussed below.

Examining examples of DA taxonomies, Jovanovic (Jovanovic et al., 2005) used the Meeting Recorder Dialogue Act (MRDA) set which was described as capable of generating millions of different labels (through combinatorial explosion). In fact their study used 7 high-level tag categories – *Statements, Questions, Backchannels and Ack., Responses, Action Motivators, Checks* and *Politeness Mechanisms*.

These categories decomposed into between one and 3 tags, for example Questions contains *Information-Request, Open-ended Question* and *Rhetorical Question*.

The taxonomy used by Fernandez contained 8 categories: *Questions, Answers, Agreement/Disagreement, Discourse Markers* (includes *Backchannels*), *Forward Functions, Control Acts, Statements* and *Other.*

The categories of *Backchannel* and *Discourse Marker* are largely used to convey information back to the speaker that the listener is understanding or paying attention (Fernandez and Picard, 2002). The categories of control acts and action motivators are

used to place an expectation of action, for example instructions and requests (but note that in both these models requests are distinct from questions).

The original MRDA set was hand-annotated using a flat structure (rather than taxonomy) of 65 DA tags; these consist largely of DAMSL tags plus some new ones such as *Floor-Grabber* and *Misspeak self-correction*. The standard MRDA tag set contains 56 DA tags, which were considered to be too challenging in terms of detail for some modeling purposes (Yang et al., 2008).

Whilst fine levels of granularity may be pertinent in QA, the work in this thesis assumes that the coarse level used by classical grammar (Quirk et al., 1985) will be appropriate to semantic similarity measurement (particularly for developing CAs) because, in essence, the variants within the coarse categories are similar in meaning. In particular, practical experience in developing scripted goal-oriented CAs leads to the belief that the most important communication takes place using questions, statements and instructions.

4.5.1.2 DA Feature extraction

The main sources of features for DA classification are surface information in the sentence text, prosodic information and contextual information. Prosodic information consists of patterns of rhythm, stress and intonation in speech and is used in telephone-based automatic speech recognition systems (Fernandez and Picard, 2002, Venkataraman et al., 2003). Statistical features have been extracted from pitch and energy including mean, range (inter-quartile), variance, skewness and kurtosis (Fernandez and Picard, 2002). Contextual information makes use of the immediate dialogue environment. Probably the most useful contextual information is the DA of the previous utterance (Keizer et al., 2002). Verbree (Verbree et al., 2006) looked further back in the dialogue by using a bigram of the last two DA labels. Traum & Hinkelman (Traum and Hinkelman, 1992) state that usually a combination of surface features and context will be necessary to disambiguate DAs. However, many applications only have access to text, so it was decided to proceed using only text surface features for classification.

Surface information used in DA Classification

One of the most common features used in a DA classifier is the **n-gram**, a string of words (typically a maximum of 4) extracted from a marked up corpus using statistical methods (Webb et al., 2005a, Venkataraman et al., 2002, Louwerse and Crossley, 2006, Ivanovic, 2005, Stolcke et al., 2000). The importance of the n-gram is that it is any n-gram may be a predictive feature for classifying the containing text. Large numbers of n-grams are

generated by any real-world corpus so a predictive subset is used. Verbree (Verbree et al., 2006) evaluated sets of 10 and 300. In n-gram terminology, a single word is a **unigram.** Unigrams with positional information have been used by (Kral et al., 2007) and Fernandez (Fernandez et al., 2008). In the Kral study the unigrams were actually unigrams of word classes (to compensate for a small training corpus) and used an (unspecified) algorithm to cluster words automatically.

The **bigram** is a particular instance of the n-gram, for example *CanYou* (Keizer et al., 2002).

Another form of n-gram which has been used is the **gappy n-gram**. These originate in string matching (Lodhi et al., 2002). A gappy n-gram is an n-gram which allows other words to occur in the sequence, thus creating gaps. Where gaps occur in a matching item it is considered to be an inexact match and therefore a penalty is applied to the match. A use in DA classification has been reported by Clark & Popescu (Clark and Popescu-Belis, 2004), who identified the 1000 most frequent words in a corpus, then used them to create the n-grams. Information about whether they were at the beginning or the end of the utterance was used to generate further features.

A combination of unigrams of the first two and last two words, plus a bigram of the first two words, was used in a study of DA classification in a meeting corpus (Liu, 2006).

Variations of the n-gram method occur in DA classification, for example **context features** and **cue phrases**.

Prasad (Prasad and Walker, 2002) used context features describing presence or absence of particular words. The attributes were Boolean in nature but concerned pairs of named entities in the utterances, for example **contains-word-RENTAL-or-CAR**.

The cue phrase, has been used in a number of studies (Webb and Liu, 2008, Verbree et al., 2006, Webb et al., 2005b). A cue phrase is a form of n-gram – a short group of contiguous words. Hirshberg defines them as linguistic expressions - such as *okay, but, now, anyway, by the way, in any case, that reminds me* – which don't convey semantics but do provide information about the structure of the dialogue (Hirschberg and Litman, 1993). For example, *now* can indicate the introduction of a new subtopic whereas *in any case* can indicate the return from a digression.

Another property of cue phrases is that they are likely to be either at the beginning or at the end of an utterance (Liu, 2006, Webb et al., 2005a, Litman, 1996). A final distinction between cue phrases and other n-grams is the fact that the first cue phrases came in the form of a specific list, the LIT set which contains 687 different phrases compiled from

various academic sources (Verbree et al., 2006) and *ipso facto* generated through human intuition.

More recently there has been an interest in automatic extraction of cue phrases from corpora (Webb and Liu, 2008), which somewhat blurs the distinction between n-grams and cue phrases. Following this method all possible n-grams are extracted (again the typical maximum length is 4 words) as candidates for the cue phrase list. Cue phrases are then selected from the list according to certain criteria. In the study by Webb the criteria were predictivity and frequency. Predictivity is predicting power of the n-gram, defined as the probability for a particular DA that an utterance containing the n-gram will be that type of DA. A cue phrase could well be a predictor of more than one DA, in which case the maximal predictivity is used, i.e. it is associated with the DA type which it predicts with the highest probability. Finally a thresholding process is used to prune the set. Those n-grams with a predictivity below a predetermined value are discarded and those with a low frequency of occurrence (in the DA they predict) are also discarded. What remains becomes the set of cue phrases.

With its focus on deeper semantic structure, LSA (Deerwester et al., 1990) forms quite a radical alternative to the n-gram approach. In this application, the features consist of a **term vector** (query) projected into the semantic space. A cosine measure is used to determine the vector of the closest matching "document" and the DA type of the match is assigned to the query (Serafin et al., 2003). Serafin's study does not, however, discuss the treatment of function words which in the past have always been removed from texts processed by LSA (Deerwester et al., 1990). A similar treatment of features was used in a Support Vector Machine study (Hacioglu and Ward, 2003), which removed stop lists and stemmed the terms to produce a vector containing the *n* most informative terms. The study also made use of LSA-like dimensional reduction.

A disadvantage of maximal predictivity as a criterion is that valuable information about the classes with lower probability is lost.

Specific examples of feature sets used in DA Classification

One approach to DA feature extraction (Andernach, 1996), used in the SCHISMA theatre booking project, was ambitious in the sophistication of the features proposed. It used a set of 7 attributes (described as cues) to classify utterances into one of 12 types. Whilst some of these types correspond to the kinds of DAs used today (e.g. greet), others are quite different (e.g. YNQ, described as having a finite verb initial position and a subject in the second position). The features consisted of:

Utterance Type: described as the mood of an utterance or syntactic category of a phrase;
Wh-word: described as the presence or absence of a Wh-word (also known as a Wh-cheft);
Subject Type: described as the type of subject of an utterance;
Cue Words: described as words which change the communicative function;
First Verb Type: described as the type of the first verb in the utterance;
Second Verb Type: described as the type of the second verb in the utterance;
Question mark: described as the presence or absence of a question mark.

It should be noted that the words for the attributes from Subject Type to Second Verb Type could be present or absent in the utterance and also that these attributes had been included on the basis of intuition (about the Dutch language). It should be noted that these features were extracted manually. Andernach felt they should be extracted by parser and was confident that PATR-II parser under development at the time would be capable of doing it. Keizer (Keizer et al., 2002) used 13 surface features to train Decision Trees (DTs) in a comparison to a Bayesian network. Features included **length,** the presence of a **Wh-word,** and the **previous DA**. There are also 3 bigram features in the starting position, **startsWithCanYou**, **startsWithCanI** and **startsWithIWant**. So if the utterance started with "Can you . . ." **startsWithCanYou** is set to TRUE and the other two **startsWith** features to FALSE. The remaining 7 features are Booleans reflecting the presence (anywhere) in the utterance of a particular lexical item. These are **Positive, Negative Okay, Tell, Do, Locative preposition** and **Locative adverb**.

A departure from normal approaches involved integrating DA classification with parsing and outputting utterances in the lambda calculus (Boye et al., 2004). Interesting as it appeared, this method relied on domain-specific pattern matching embedded in the parser. Verbree (Verbree et al., 2006) used some quite elaborate features alongside the more conventional ones and compared results across 3 corpora. The unusual features included **n-grams of POS-tags** and **number of occurrences of the word OR in the presence of a question-mark**. Also n-grams were allowed to be non-order specific and a compression technique was used to reduce the large amount of data generated by merely recording the presence for each of a large number of n-gram values. Unfortunately the study discovered that different combinations of the features were required to obtain the optimal performance in each of the different domains. Parser labels were also used in a bilingual study by Levin (Levin et al., 2003). The English part of the study made use of 212 features indicating the presence or absence of a particular DA.

A related study (Litman, 1996), which does not classify DAs, but rather classifies Cue Phrases into one of two types – Discourse or Sentential, made use of a complex collection of features including a prosodic set, class information about previous cue phrases, **orthographic** features (e.g. comma – but strangely does not mention question marks), **part-of-speech** features (e.g. conjunction) and **tokenised lexical** features. There were 35 words in the lexical feature set which appear to have been chosen on a fairly ad-hoc basis. There are number of adverbs e.g. *Actually*, some function words e.g. *And* (co-ordinating conjunction) and some words from multiple classes e.g. *Like* which belong to multiple classes (verb, noun, preposition).

A final thought is that some aspects of feature selection are likely to be language dependent.

Discussion of Feature extraction

Andernach's approach initially appeared relevant to this study and was used with a Kohonen network (Andernach et al., 1997). His experiment used manual feature extraction and no automatic method of extracting them has been reported. Also, a search of the literature revealed no further use in the intervening years and different approaches being adopted by his colleagues (Kral et al., 2006, Kral et al., 2007).

N-grams are the most established feature used for DA classification. This may be related to the strong tradition of using statistical processes and Bayesian techniques in NLP. Choice of n-grams assumes independence between them and simply selects those with the highest probability of predicting particular DAs (discussed previously in Cue Phrases). This could contribute to misclassification of utterances from all of the other DAs the n-gram occurs in (but with lower probability than the predicted DA). Additionally, longer n-grams have highly specific targeting power, but at the expense of generalisation. It may be argued that gappy n-grams have greater generalising power, but because they penalise the gaps they could perform poorly when the gap actually makes a positive contribution to the feature. Also trigrams and quadragrams have potential robustness problems in the presence of speech recognition errors in systems with audio input (leading to false negatives). Therefore it is worth investigating methods that can discover complex patterns of features distributed throughout the user utterance.

Use of the question mark raises interesting problems. If the question mark is used perfectly there is no need for any other feature and the task becomes trivial. On the other hand, if errors are made in using such a conclusive feature, significant noise may be introduced into the system.

Feature extraction in prior work leaves much to be desired. The ambitious set used by Andernach has proved infeasible for automated extractions. Furthermore, even if these could be made to work they would require a computationally intensive chain of processes such as parsing. The basic n-gram systems use a statistical selection method which fails to exploit complex relationships between attributes in forming features. Gappy n-grams do not improve upon this position. The requirement is for a feature set which is simple to extract and computationally efficient.

Function word features

Previous practical experience in producing pattern matching CAs revealed that words such as *who, to, me, did* and *not* (and their positions in the sentence) were doing a lot of the work involved in discriminating between different types of sentence about the same subject matter, particularly in identifying questions. These words belong to a closed class known as "function words."

Chapter 3 of this thesis described the division of the English language into Content and Function word classes then concentrated on content words. This is how semantic similarity has always worked to date, because content words convey the propositional content. According to LSA, for example, function words have no role whatsoever in semantic similarity measurement. The hypothesis explored by this chapter is that in conveying the structural information about a sentence, function words alone will provide sufficient information to classify DAs.

Content words are members of open classes, that is new words and meanings are invented on a continuous basis. Consequently algorithms that use them can never be completely up-to-date. In contrast, there are a relatively small number of function words and they are fixed; the English language has evolved all that it needs. Therefore algorithms that use function words will be stable over time.

Despite the fact that function words are a closed class, a definitive list is not easy to come by. This is because researchers in Information Retrieval (IR), the primary field of use, are more interested in an associated class, stop words. Stop word lists contain high frequency words. Consequently they contain most of the function words plus some high frequency content words. However they omit low frequency function words.

Therefore, for the work in this thesis, a list of function words was compiled by merging the available stop lists, removing the content words and adding low frequency function words through dictionary searching. The low frequency function words tend to consist of legalistic terms such as *whereafter*.

The potential of function word features can be illustrated by an example. Consider the sentence "Did you pay on time?" The words and their positions make up a set of attributes of the sentence. This question can be recognised from the fact that the first word position contains *did* and the second position contains *you*. Both are function words: *Did* is an auxiliary verb and *you* is a pronoun. The combination of an auxiliary verb in the first word position and a pronoun in the second could be a good combination of features to indicate a question.

Consequently it was decided to use a combination of function word type and position as the features for DA classification. This will be explained in 4.7.6.

4.5.1.3 Dialogue Act Classifiers

Despite the fact that DA classification is not a large field of study, an interesting range of classifiers have been used in studies to date. These include statistical (Lesch et al., 2005) and n-gram models (Webb et al., 2005a), Bayesian networks (Keizer et al., 2002), Naïve Bayesian Classifiers (Kral et al., 2007), Kohonen Networks (Andernach et al., 1997), Multi-Layer Perceptron (Kral et al., 2007), Backpropagation Artificial Neural Networks (Levin et al., 2003), Maximum Entropy (Clark and Popescu-Belis, 2004), C4.5 Decision Trees (Verbree et al., 2006), Production Rules (Prasad and Walker, 2002), Simple Heuristics (Webb et al., 2005b), Hidden Markov Models (Stolcke et al., 2000), Partially-Observable Markov (Bui et al., 2007), K-Nearest Neighbour (Levin et al., 2003), Support Vector Machines (Fernandez and Picard, 2002), Learning Vector Quantisation (Jokinen et al., 2001) and Self Organising Maps (Jokinen et al., 2001). It is difficult to draw any conclusions about the relative performances of particular classifiers, because different studies use different features, different DA taxonomies and different corpora – so results are not comparable.

It is however, possible to observe that Bayesian techniques are popular, with a number of studies using Bayesian Networks (Keizer et al., 2002, Andernach, 1996, Bilmes, 2005, Bui et al., 2007, Jovanovic and op den Akker, 2004) and Naïve Bayesian classifiers (Kral et al., 2007, Levin et al., 2003, Ivanovic, 2005, Wiebe et al., 1999).

Lori Levin's (Levin et al., 2003) study compared the naïve Bayesian classifier, k-Nearest-Neighbour, decision tree and neural network methods. The Bayesian network performed quite poorly, initially, but improved somewhat when it was trained on simple word bigrams (as opposed to the unigram model over arguments). However, when combinations of classifiers were evaluated the pairing of the k-Nearest-

Neighbour and Bayesian classifiers were most effective (81.25% Classification Accuracy on English and 78.93% on German).

4.5.2 Selection of a classifier for Dialogue Acts

Implementation of the new model of STSS requires a DA classifier which is both accurate in classification and computationally efficient. It would also be useful if the classifier were transparent, i.e. it were possible to understand how it reached its classification decisions. In preparation for the full methodology, which is described in section 4.7 , an initial experiment (experiment 4.0) was conducted to investigate these issues. Experiment 4.0 used a preliminary dataset which was intended to be challenging, mixing straightforward and difficult non-questions together. It was not completely balanced, consisting of 591 questions and 615 non-questions. This size of dataset, used with 60-fold cross validation gives approximately 1,180 training cases and approximately 20 testing cases for each fold. These proportions were chosen based on a rule of thumb for training set size proposed by Quinlan (Quinlan, 1993).

The preliminary dataset was not used again in this chapter, but it was found to be useful in chapter 5 where unbalanced datasets were the norm.

Four well-established classifiers were tested: Naïve Bayesian, Multi-Layer Perceptron, Bayesian Network and C4.5 DT. The motivation was to get a quick view of which classifier was most promising for the detailed work that follows. Consequently the experiment was conducted using the default parameter settings in the Weka package (Witten and Eibe, 2005) and other conditions were held the same as those in the experiments described later in this chapter. The results are shown in table 4.7.

Classifier	Naïve Bayes	MLP	Bayes Net	C4.5
CA	55.98	69.14	77.87	88.73

Table 4.7 Classification accuracies for popular classifiers

The C4.5 decision tree performed best. Whilst this is not conclusive certain other factors favoured decision trees:

- The rule induction method of decision tree construction is both well understood and tried-and-tested.
- Decision trees are transparent: it is possible to see how the decision tree has made its classification.
- Decision trees are scalable to large numbers of instances running simultaneously.

The last factor is highly important in information-providing web applications. For example the UK national flu service received 9.3 million hits per hour on the first day of operation (BBC, 2009).

Consequently, it was decided to use DTs to implement the DA classifiers in chapters 4 and 5 of this thesis. The next section provides a detailed analysis of decision trees to inform parameter setting and the encoding of features from sentences.

4.6 Decision Trees

Decision Tree (DT) induction is a highly effective method of machine learning for classification. The experiments reported in this chapter make use of decision trees constructed using the C4.5 algorithm as implemented in the Weka package (Witten and Eibe, 2005). C4.5, one of the most popular decision tree algorithms, is a descendant of the highly influential ID3 algorithm, both of which were created by J. Ross Quinlan (Quinlan, 1986, Quinlan, 1993). ID3 itself was inspired by Concept Learning Systems developed two decades earlier (Hunt et al., 1966).

DTs partition the input space recursively, through a hierarchy of tests and the outcome is a set of rules induced from a training set of example instances previously labelled with their classes. When a new case is supplied and the tree is traversed to completion, the result is the classification of the input case into one of a number of categories or classes. To do this it is supplied with the values of a number of attributes of the entity to be classified. One important feature of C4.5 is that, although it can take both continuous and discrete attributes as inputs, it is restricted to classification of input cases to one of a small number of discrete classes (usually 2). Alternative forms of tree exist which can predict the value of a continuous output variable from input attributes, these are described as "regression" trees, for example CART (Breiman et al., 1984).

The work in this thesis follows the most common approach of using two classes (questions and non-questions). One method of constructing DTs is to use a process called knowledge engineering to extract rules (which become tests in the DT) from the knowledge of one or more human experts. This can be problematic for a number of reasons. Sometimes humans are not able to explain how they analyse the data and take decisions (Michie, 1993). Sometimes they will give post-hoc rationalisations of a process they do not understand. Occasionally human experts will come to the same conclusion in a classification task but will disagree on how they interpreted the evidence to get there.

The alternative is to use an automatic technique to induce a DT from a set of cases (the training set). Later this tree can be validated using cases it has never seen before (the test set), then deployed in a real-world application (for example mortgage application processing).

The starting point of a DT is one node with all examples labelled as belonging to one class. A node is defined as 'impure' when all the examples reaching that node are not in the same class as the node itself. During training, successive partitioning leads to increasingly pure nodes in the tree. Left to itself, an algorithm such as C4.5 would continue to add decision nodes, splitting the data until all of the leaves were pure (contain a single class) or the attributes have been tested exhaustively. If the leaves are pure they are labelled with the appropriate class, if they are impure they are labelled with the class having the largest number of instances at the leaf.

4.6.1 Decision Tree construction

The process of constructing a tree (as described by Quinlan) is a "divide and conquer process". Given a training set T containing classes {C1,C2,…Cn} there are three possible actions:

- If T is pure (contains one or more cases of the same class), the DT for T is a leaf labelled with the contained class.

- If T is empty (contains no cases), the DT is a leaf whose label (class) must be inferred from other information.

- If T contains cases belonging to more than one class, select an appropriate test which partitions T into a number of subsets which are progressively purer. The test must operate on a single attribute, produce mutually exclusive outcomes and cover all the cases in T. This produces a node in the tree at which branching takes place (as a result of the test).

The process continues along all of the generated branches of the tree until a set of pure leaves is generated (i.e. each leaf contains only one class). However, the generation of a useful tree does not end in this manner. Such a tree may classify the training set perfectly but it is likely to be over-trained (to have memorised the training set rather than modelling the principles of the domain). Therefore trees undergo a pruning process, which makes them better at generalisation but results in at least some of the leaves being impure.

One of the most important factors in constructing a DT is the selection of a test to split the current set of training cases into partitions. This test involves choosing the attribute to split on and if it is continuous, the point in the range to split at.

The criterion used in C4.5 is known as "gain ratio" which is a development of "information gain" used in ID3. The basic principle behind information gain is that allocating a training case to class is effectively sending a message about the case and consequently there is an average (over all the cases) amount of information to identify the class of a case. Partitioning into smaller sets increases the average amount of information and the partitioning rule which maximises the information gain is the one selected to build the tree.

For a particular training case, from a set of S messages which belongs to a class C_j:

$$\inf[S] = -\sum_{j=1}^{k} \frac{freq[C_j, S]}{|S|} \times \log_2\left(\frac{freq[C_j, S]}{|S|}\right) \text{ bits} \qquad (4.3)$$

Suppose previous training steps have left a set T and a test X partitions this into n subsets, then:

$$\inf_x[T] = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \inf[T_i] \qquad (4.4)$$

And information gain is calculated as:

$$gain = \inf[T] - \inf_x[T] \qquad (4.5)$$

The gain ratio criterion for test selection was developed to overcome a problem of the information gain test that it was biased in favour of tests with many outcomes. This uses a factor *split information* to normalise the gain resulting in the gain ratio:

$$splitin[X] = -\sum_{i=1}^{n} \frac{|T_i|}{T} \times \log_2\left(\frac{|T_i|}{|T|}\right) \qquad (4.6)$$

And

$$gain_ratio(X) = gain(X) / splitin(X) \quad (4.7)$$

Quinlan notes that gain ratio may be biased against attributes with many values.
When searching for the splitting point for a continuous attribute, C4.5 sorts all of the values that occur in the training set into ascending order and evaluates all of the midpoints between existing pairs as potential split points. When one is found, it uses the largest existing value in the training set below this midpoint as the actual split point (the rationale for this seems to be that a DT that uses values which actually occur in the training set will be clearer to read).

4.6.2 Decision Tree pruning

In the worst case, a tree that was optimised for the training set could grow to a point where each leaf represented a single case from the training set. Such a tree has completely memorised the training set and would be a very poor generaliser when presented with new data to classify.
Pruning involves removing nodes from the tree to make it generalise more effectively. There are two possible approaches to pruning, restricting the growth of the tree during construction (prepruning) or removal of parts of the tree after construction (postpruning). In particular, postpruning (backward pruning) for subtree replacement involves removing a subtree and replacing it by a single leaf.
C4.5 uses the postpruning method. It builds the complete tree before pruning, on the grounds that the drawback of increased computation time is offset by the fact that the possible partitioning is explored more thoroughly.
In C4.5 the decision to prune at a particular node is based on comparing the expected error rate for the original subtree with that for the replacement node. "Expected error rate" refers to the error rate that would be expected if the tree were run with an independently selected test dataset.
An error occurs when an instance presented to the tree reaches a leaf and is not of the labeled class. If there are E errors out of N instances classified at the leaf, the measured error rate, f, is

$$f = \frac{E}{N} \quad (4.8)$$

Simply using an error calculated from the training data would not be effective because the tree is already optimised for that set. When large amounts of data are available, Reduced Error pruning is feasible. This requires a validation set which can be used for postpruning. This is not feasible where there is limited data (typically cases where n-fold cross-validation will be used). Here an error rate is calculated from the training data; however it is modified to produce a pessimistic estimate of the error rate for use in the calculations. Given an observed error rate, 25% confidence intervals are calculated for this rate and the upper bound of the confidence interval is substituted in the pruning calculation. Given a true error rate, q, and a desired confidence level, c, the confidence limits, z may be calculated from equation 4.9:

$$\Pr\left[\frac{f-q}{\sqrt{q(1-q)/N}} > z\right] = c \qquad (4.9)$$

and the pessimistic estimate for the error rate e is the upper limit expressed by equation 4.10:

$$e = \frac{f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}} \qquad (4.10)$$

In C4.5 the default confidence setting is 25% which leads to a value of 0.69 (measured in standard deviations) for z. Although Quinlan has supplied a comprehensive specification for C4.5 in the literature (Quinlan, 1993) there is no justification for choosing that particular value, other than the statement "the default value seems to work reasonably well for many tasks." On various occasions in describing C4.5 Quinlan admits to taking a fairly pragmatic approach in adopting techniques. In justifying this pruning technique he states (Quinlan, 1993):

"Now, this description does violence to statistical notions of sampling and confidence limits, so the reasoning should be taken with a large grain of salt. Like many heuristics with questionable underpinnings, however, the estimates that it produces seem frequently to yield acceptable results."

Although confidence level pruning works well, there is a minor practical problem with its use. The objective with pruning is to find the most heavily pruned tree that has not fallen

significantly in CA below the benchmark. To establish this, pruning must be driven to the point where this fall occurs and then find the tree with a CA just above that point. With some datasets the confidence level is reduced to a number so small that the algorithm crashes (probably a floating point underflow) before the significant fall in CA occurs. In these cases the trees with the lowest confidence levels achieved before the crash are reported.

An alternative pruning method for C4.5 is the minimum weight, m, parameter implemented in the Weka toolkit as MinNumObjs (Witten and Eibe, 2005). This method, allows the user to specify the minimum number of instances which can occur at a leaf. This was originally introduced to deal with some near-trivial tests in which almost all of the cases have the same outcome, so one branch would represent the majority of the cases and the other very few. By setting the minimum number of cases that can occur in a partition this can be prevented. During pruning subtrees are removed and their instances re-distributed amongst higher nodes which have themselves become leaves. The default minimum value is 2 but higher values can be used as a pruning strategy and Quinlan (Quinlan, 1993) suggests it as technique where there is significant noise in the data. MinNumObjs (MNO) pruning is not subject to the arithmetic problems of confidence level pruning and always finds a point at which the CA drops significantly below that for the default value.

Another interesting observation was that with some datasets there was an initial significant drop in CA, followed by a recovery as pruning continued, followed by another, consistent drop in CA. This did not result in CAs that significantly changed the situation, so the results around the initial drop are reported in the following experiments.

One of the chief advantages of DTs over other AI techniques such as Artificial Neural Networks (ANNs) is that through the combination of pruning and the production of rules it is possible for them to "explain" how they reach their decisions and thus provide us with a greater insight into the problem domain. However, this in itself depends on selection and encoding of the features set.

Several other properties of DTs that have a bearing on the following series of experiments are discussed in the following sections.

4.6.3 Coping with missing attribute values

The first is the ability to accommodate datasets where the cases have missing values. There are a number of ways of dealing with this including using some inferencing techniques to fill in the missing values, for example simply by substituting the most frequent value for

the attribute. However C4.5 uses a more sophisticated approach. Where an attribute value is unknown, a weight is calculated for each possible attribute value based on the probability of the particular value occurring for the attribute. The weights are summed for the relevant branches and the values propagated to the leaves, which can contain fractional numbers of cases. A consequence of this is that MNO pruning discussed above is a little more subtle. It is not the minimum number of objects in a leaf but the minimum value of the sum of the weights of the cases in the leaf.

4.6.4 Windowing

Windowing was introduced as part of the ID3 DT construction algorithm when the Random Access Memory available to computers was restricted. It has been retained in C4.5 for a number of reasons, including speed of tree construction and the building of better trees through selection of an initial window which has a better distribution of data than the training set as a whole. It involves selecting a subset of the training data (the window) to train an initial tree, classifying the full training set, then selecting a proportion of the misclassified cases from the training set to add to the window. It is an iterative process and continues until all of the training set is correctly satisfied or a halting condition (for coping with noisy data) is matched. Because there is an element of randomisation, this is one of the reasons why *different* DTs can be generated from the *same* training set and with the *same* parameters.

4.6.5 Grouped attribute values

Grouping involves combining certain values of an attribute and treating them as a single value. The example given in (Quinlan, 1993) uses the soybean dataset where one attribute *canker value* has the values *none, below soil, above soil* and *above* 2nd *node*. The gain ratio test is used to decide whether or not to merge values into a group and in this case merging *above soil* and *above* 2nd *node* is found to be favourable.

4.6.6 Cross Validation

C4.5 includes code to support the use of cross-validation. This involves randomly splitting all of the available data into a number of equally-sized folds (the default is 10 folds). Rotating through the set of folds, one is chosen as the test data and the rest combined to form the training data for that fold. Consequently, all of the data can take part in training and testing, without any particular DT having seen its test data during training. Based on comments about training set size made by Quinlan (Quinlan, 1993), it was decided to use

over 1,000 training cases in the experiments in this chapter, by using 600 questions, 600 non-questions and using 60-fold cross validation.

4.6.7 Illustrative examples of Decision Tree construction

The following small example is intended to serve 3 purposes:

- to illustrate the principles of decision tree construction
- to show how function words may be used as features
- to show how the grouping approach can be used to produce better trees

Table 4.8 lists 16 short texts (extracted from the dataset described in section 4.7); each short text is spread over the first three columns. The first word (W1) is in the first column, the second word (W2) is in the second column and the remainder of the short text (Remainder) is in column 3. Note that all of the words in columns 1 and 2 are function words.

The final column contains the classification, Q = question, N = non-question. The texts are derived from real-word FAQ lists and blogs found on the World Wide Web and show that the first two words contain important features for classifying the question DA.

W1	W2	Remainder	Class
We	are	all about the city	N
What	a	vulgar, self-publicising woman she is	N
What	a	wonderful end to the day: the hunter was home from the hill	N
What	are	fuel cells	Q
What	are	some of the false and real causes of male hair loss	Q
What	can	I do for a color change	Q
What	did	you do when your card was stolen	Q
What	did	you achieve in your last role	Q
When	are	the quarterly estimated tax returns due	Q
When	I	get a spare moment all I can think about is my warm, cosy bed	N
When	I	went to college, he was still in high school	N
Where	a	parent might not be sometimes a teacher will be listened to	N
Where	are	the enlarged group's headquarters	Q

Where	can	I find an answer to a query I have about a cbbc programme	Q
Where	I	would really like to go is Urbino in Italy	N
You	did	Run faster when you were younger	N

Table 4.8 Set of training cases, T, containing 8 questions and 8 non-questions

Suppose that, following a process of feature extraction, W1, W2 and Class are presented to a DT construction program; how will the tests be selected to build the tree?

First of all, the information content of T is:

Info(T) = $-8/16 \times \log_2(8/16) - 8/16 \times \log_2(8/16) = -0.5 \times -1.0 + -0.5 \times -1.0 = +1.0$

There are two discrete attributes so each will be evaluated for a multi-way split of the training set. As described above, C4.5 compares the gain ratio measures for each of the tests and chooses the test with the highest value. Therefore it is necessary to calculate this for tests on both the attributes.

Gain Ratio for W1

After using the attribute W1 to divide T into 5 subsets (We, What, When, Where, You) the new information content is:

Infox = 1/16 x (-0/1 x log2(0/1) -1/1 x log2(1/1) we
+ 7/16 x (-5/7 x log2(5/7) – 2/7 x log2(2/7) what
+ 3/16 x (-1/3 x log2(1/2) – 2/3 x log2(2/3) when
+ 4/16 x (-2/4 x log2(2/4) – 2/4 x log2(2/4) where
+ 1/16 x (-1/16 x log2(0/16) – 1/16 x log2(1/16) you

infox = 0.79980 bits

Information gain from split on W1 = 0.2002

Split info = -1/16x log2(1/16) - 7/16 x log2(7/6) - 3/16 x log2(3/16) - 4/14 x log2(4/16) - 1/16 x log2(1/16)
= 1.97 bits

gain ratio = 0.10

Gain Ratio for W2

After using W2 as the attribute to divide T into 5 subsets (a, are, can, did, I) the new information content is:

Infox= 3/16 x (-0/3 x log2(0/3) – 3/3 x log2(3/3)	a
+ 5/16 x (-4/5 x log2(4/5) – 1/5 x log2(1/5)	are
+ 2/16 x (-2/2 x log2(2/2) – 0/2xlog2(0/2)	can
+ 3/16 x (-2/3 x log2(2/3) – 1/3 x log2(1/3)	did
+ 3/16 x (-0/3 x log2(0/3) – 3/3 x log2(3/3)	I

infox= 0.15360 bits

Information gain from split on W2 = 0.85

Splitinfo = -3/16 x log2(3/16) - 5/16 x log2(5/16) – 2/16 x log2(2/16) – 3/16 x lo2(3/16) – 3/16 x lo2(3/16)
= 2.26 bits

gain ratio = 0.37

Therefore the test on W2 is chosen over W1 because of its greater gain ratio. It is also worth observing that the older test of simply using information gain would achieve the same result.

Training a C4.5 decision tree (using the C4.5 version 8 classifier in the Weka Toolkit) shows that W2 is indeed selected for the first split. Using the default confidence level of 0.25 for pruning, the tree is restricted to the split on W2 and a CA of 87.5 is achieved. Changing confidence to 0.75, the tree grows to the point where it achieves a CA of 100. The trees are shown in table 4.9. This small example has brought out a problem with DT construction. The gain ratio calculations for data revealed that both attributes (W1 and W2) had something interesting to say about classifying the cases. The problem is that after the split on W2, the W1 information has been diluted in the 5 subsets created and can not be

exploited effectively. The pruned tree misclassifies 2 cases and the larger tree is overtrained (contains too much detail and would not generalise well to new data).

C4.5 confidence=0.25	C4.5 confidence=0.75
W2 = a: N (3.0) W2 = are: Q (5.0/1.0) W2 = can: Q (2.0) W2 = did: Q (3.0/1.0) W2 = I: N (3.0)	W2 = a: N (3.0) W2 = are | W1 = What: Q (2.0) | W1 = Where: Q (1.0) | W1 = We: N (1.0) | W1 = When: Q (1.0) | W1 = You: Q (0.0) W2 = can: Q (2.0) W2 = did | W1 = What: Q (2.0) | W1 = Where: Q (0.0) | W1 = We: Q (0.0) | W1 = When: Q (0.0) | W1 = You: N (1.0) W2 = I: N (3.0)

Table 4.9 Decision tree structures for the two C4.5 classifiers

Quinlan introduced the grouping algorithm to solve this problem. Suppose we replace the words in W1 and W2 by more sophisticated information about the word's functional class. This results in table 4.10:

W1	W2	Class
P	V	N
P	V	N
Wc	A	N
Wc	A	N
Wc	A	N
Wc	P	N
Wc	P	N
Wc	P	N
Wc	V	Q
Wc	V	Q
Wc	V	Q
Wc	V	Q

Wc	V	Q
Wc	V	Q
Wc	V	Q
Wc	V	Q

Table 4.10 Attribute values grouped by function

Where P = pronoun (I, You, We etc.), Wc = Wh-cheft (What, When, Where etc.), V = auxiliary verb (Are, Can, Did etc.) and A = article (a, the).
Using the default parameters confidence=0.25 and MNO=2, the result is a DT with 6 nodes (4 of which are leaves) and 100% Classification Accuracy on the training data:

W2 = A: N (3.0)
W2 = V
| W1 = Wc: Q (8.0)
| W1 = P: N (2.0)
W2 = P: N (3.0)

Grouping is an important consideration. It can be applied when designing the feature extraction process, however identifying more sophisticated features is non-trivial in real-world applications.

4.7 The proposed Slim Function Word Classifier algorithm

The proposed algorithm will classify a dialogue act, against a background of other dialogue acts. Applications requiring the classification could include Conversational Agents or Information Retrieval systems. The algorithm is composed of the following steps:

1. The algorithm receives a sentence of up to 25 words from the application requiring the classification
2. The sentence is pre-processed
    a. All punctuation apart from apostrophes is stripped from the utterance
    b. All contractions are expanded (e.g. *won't* is transformed to *will not*)
    c. All extra spaces between words are stripped from the utterance
3. The utterance string is processed word-by word to produce a tokenised representation. For each word in the utterance

if the word exists in the function word table

the corresponding token is inserted into the tokenised string

else

the wildcard token is inserted into the tokenised string

endif

4. For each remaining empty slot in the tokenised string

   Fill the empty slot with word-not-present token
5. The record is submitted to a trained DT for classification
6. The classification returned by the DT is passed to application requiring the classification

The rest of this section describes the experimental methodology for creating decision trees and the method of pre-processing the data. The experimental methodology involved:

- analysing the various types of questions and non-questions that might require classification
- formulating hypotheses about the performance of DA classifiers for combinations of these questions and non-questions
- collection of raw question / non-question data
- creating datasets for particular combinations of the question / non-question forms
- preprocessing the datasets for feature extraction
- constructing DT classifiers for specific hypotheses
- testing the DT classifiers, particularly with respect to pruning and generalisation

4.7.1 Analysis of questions and non-questions

Due to the size and complexity of the English language (discussed in chapter 3) there is a problem of collecting a representative sample of questions and non-questions.

This section considers two views of question classification:

- Grammatical classification (Quirk et al., 1985)
- Domain-based classification (Flynn, 2002b) (Christensen and Garvin, 1992)

The highest level of grammatical distinction between different types of sentences is between Simple and Multiple sentences.

4.7.1.1 Grammatical classification

Classes of questions composed of Simple Sentences

A Simple Sentence contains a single independent clause. The standard grammatical classes of questions derived from simple sentences, with one or two examples in each case (Quirk et al., 1985) are listed below:

1. Yes-No
   Have you finished the book?
   Don't you believe me?
2. Wh-questions
   What is your name?
   Who hasn't had any coffee?
3. Option selection
   Would you like to go for a walk or stay at home?
4. Tag questions
   They forgot to attend the lecture, am I right?
   The boat hasn't left, has it?
5. Declarative questions*
   You've got the explosive?
   You are leaving already?
6. Exclamatory
   Hasn't she grown?
7. Rhetorical
   Who cares?
   Am I alone in thinking the motion is out of order?

* Without prosodic information the declarative question form would be interpreted as a declarative rather than a question. Declarative, exclamatory and rhetorical are all dependent on prosodic information for their correct interpretation. This can be resolved using the question mark. As Quirk observed (Quirk et al., 1985) "… the question mark matches in writing the prosodic contrast between this sentence as a question and the same sentence as a statement." The exclamatory and rhetorical categories do not implicitly ask for information; uses include emphasis or giving an opinion. Given the text-based approach, declarative, exclamatory and rhetorical questions are not suitable for inclusion in this study.

Classes of questions composed of Multiple Sentences

Grammar recognises two more sophisticated question categories which could be more challenging to the techniques developed in this work – these are questions embedded in Compound Sentences and Complex Sentences.

A Compound sentence contains two or more co-ordinated main clauses, for example:

I admire her reasoning but I reject her conclusions.
What is your name and where do you live?

A complex sentence has a single main clause and one or more subordinate clauses, for example:

Although I admire her reasoning, I reject her conclusions.
Can you confirm which flight we are taking?

The complex form allows questions to participate in indirect DAs, for example:

Please confirm which flight we are taking.

is a directive (instruction) which seeks the same information as the question *Which flight are we taking?*

4.7.1.2 Domain-based classification

An example of domain-specific categorisation is the assessment of medical students. A list of categories, with examples, derived from Christensen (Christensen and Garvin, 1992) is shown below:

1. Information-Seeking Questions:

What were the blood values from the lab?

2. Diagnostic Questions:

What conclusions did you draw from these data?

3. Challenge (Testing) Questions:

What evidence supports your conclusion?

4. Hypothetical Questions:

If the liver function tests were normal, how would that have affected your treatment plan?

5. Action Questions:

What needs to be done to implement the plan for this patient?

6. Extension Questions:

What are the implications of your conclusions for the treatment of asthma among children in elementary school in our community?

7. Priority/Sequence Questions:

Given the patient's limited resources, what is the first step to be taken?

8. Prediction Questions:

If your plan (conclusion) is appropriate, what do you expect to happen over the next month? Year?

9. Generalisation Questions:

Based on your experience and the studies of the incidence of teenage pregnancy, what do you consider to be the most effective strategies for our local high school teachers and counsellors?

Of the above categories, 1,2,3,5 and 6 are questions comprised of simple sentences and 4, 7, 8 and 9 are complex questions. From the point of view of constructing a dataset for this study, there is nothing to be gained from discriminating between questions such as *What were the blood values from the lab?* and *What evidence supports your conclusion?* Some other types of question were found during the investigation:

- indirect questions:
  Joan asked was he ready yet.
- directives containing questions:
  Please confirm which flight we are taking.
- assertives implying questions:
  I have received an invoice and I do not know what it relates to.
- leading questions:
  You did have the gun when you left, didn't you?

The leading question is a specialist form from the legal domain and is easily disposed of. Its distinguishing feature is that it implies a correct (or known) answer. Structurally these questions will fall into one of the categories described previously and therefore will be

covered automatically (the example given is a negative form of tag question). The indirect questions don't necessarily increase the complexity of the recognition task over and above the existing categories. The problem is deciding what to do in response – how they would affect a Conversational Agent's intentionality within its task and problem domain. This is beyond the scope of the present work so this question form will not be included.
The Directive corresponds to W*hich flight are we taking?* and the Assertive corresponds to *Can you explain what the invoice I have received relates to?*
In fact, both are indirect dialogue acts; they seek information but are prima facie, non-question. As such they will not be included in the current work (although the examples look promising in terms of the function words contained in the sentences).

4.7.2 Forms of questions used in this study

4.7.2.1 Straightforward questions

The Straightforward question type is, effectively the question contained in a Simple Sentence. The term "Straightforward" has been adopted to indicate that although the question itself may be quite sophisticated or difficult to answer, the form of the question is likely to be the least challenging for a DA classifier.
These questions are short and to the point. They do not require resolution of references to prior dialogue and they do not contain pertinent information embedded in clauses separated from the main question clause. Good examples of this simplest form, such as "When was James Dean born?" can be found in the TREC factoid set. Such questions have a very obvious feature in the first word position – the presence of a Wh-cheft.
Straightforward questions cover the grammatical classes Yes-No, Wh-questions and Option selection from the simple sentences class and provide evidence for some questions from the multiple sentences class. For example, they will provide evidence that the first question in a compound question will be recognised. They also cover the domain-based classes Information-Seeking, Diagnostic, Challenge, Action and Extension.

4.7.2.2 Straightforward questions with preambles

It is possible to create a more difficult class of questions by shifting the first word (typically an interrogative introducer or wh-word) of the question further down the sentence. This can occur in a form known as the "Pushdown" where it moves part of another clause into the main clause e.g. *On what side of the road was he driving?*

Introductory words and phrases, which do not qualify as clauses in their own right, can also be used with a question. They have a variety of purposes which don't actually contribute to the semantics of the surface, for example for continuity, politeness, attention grabbing etc.

These words and phrases are accommodated for in this study by adding a "preamble" of a few words to the start of the sentence. In distinguishing between a preamble and a clause the preamble was limited to a maximum of 5 words and specifically does not contribute to the semantic content of the utterance. Thus examples include:

Actually
Almost everyone asked
And there is another thing

So a question of this kind would be

And there is another thing, when was James Dean born?"

These preambles look remarkably similar to Cue Phrases, however the two properties of a good preamble are that (i) it must "work" when placed in front of a question (like a cue phrase) and (ii) it must not provide evidence to the classifier that it is part of a question (contrary to a cue phrase). The middle phrase, *Almost everyone asked*, may appear to contradict these requirements. However, the key evidence for a question is in the word *asked*, a verb which will not be considered by the SFWC. When producing preambles for non-questions it was fairly straightforward to replace preambles of this type with equivalents that worked for non-questions but generated the same tokens. For example, *Almost everyone denied.*

Straight forward questions with preambles cover short Tag questions from the grammatical class and short versions of the Hypothetical, Priority/Sequence Prediction Questions and Generalisation questions from the domain-based classes.

4.7.2.3 Simulated clauses

The need to represent the grammatical category of Questions in Multiple Sentences posed some problems. One source of data, the IRS FAQ set (IRS, 2009) contained some questions fitting this form:

For business travel, are there limits on the amounts deductible for meals?

If I claim my daughter as a dependent because she is a full-time college student, can she claim herself as a dependent when she files her return?

If you pay child support, are you allowed to deduct anything on your taxes or claim the child as an exemption?

I received a Form 1099-MISC instead of a Form W-2. I'm not self-employed, I do not have a business. How do I report this income?

The worst form of the last form of question is one that tends to occur in computer helpdesk applications, which contains a relatively long description of the current state of a computer followed by "…Can you help?"

Preliminary work revealed that attempting to collect examples of this data from real-world dialogue sources would be very heavily time consuming and have a low productivity. However, it was also clear that for this category, the questions were again straightforward questions with one or more clauses prefixing them. Therefore, for this set of experiments it was decided to prefix questions with non-questions to simulate the clauses. This may provide a harder classification task because there are no relationships between the function words in the non-questions and the questions that they prefix.

The IRS source is particularly interesting, because much of the previously reviewed work on DA classification concerned itself with problems involved with spoken dialogue such as dealing with incomplete utterances and channel management. However, when these are resolved the dialogue content is relatively simple. The IRS site reveals that even when these modal problems of dialogue have been resolved, the underlying problems of communicating in a real-world goal-oriented system are rich and complex.

Simulated clauses cover longer Tag and Multiple Sentences questions from the grammatical class, in particular the second question of a compound question.

They also cover longer versions of the Hypothetical, Priority/Sequence Prediction Questions and Generalisation questions from the domain-based classes.

4.7.2.4 Omitted question classes

The grammatical class Declarative was omitted because the only way it can be identified is with prosodic information. Prosodic information is beyond the scope of this thesis, which deals with text-based input. However, the literature reviewed in section 4.5.1.2 suggests that this information could be used in future DT classifiers.

Some questions composed of multiple sentences have also not been covered. In particular, a question in an Indirect DA could confuse a single DA classifier. However, it is anticipated that multi-classifiers could be produced in future, for example the sentence:

Please confirm which flight we are taking.

could be classified as Instruction containing Question.

4.7.2.5 Straightforward vs. Difficult non-questions

The problem of "Difficult non-questions" arises from testing of early pattern matching systems developed by the MMU Centre for Conversational Agents conducted by Donald Michie. Simple patterns based on the occurrence of Wh-chefts break down when non-questions are framed using the Wh-chefts as pronouns or conjunctions. In particular, it is possible to construct valid non-questions with pronoun Wh-chefts or conjunctions as starting words. An example taken from a hair / beauty blog site (http://www.kaboodle.com/reviews/psoriasisnet-6) is:

When psoriasis develops on the scalp, hair loss sometimes follows.

and this needs to be represented in training and testing data used in the following experiments.

The term "difficult" is used because although sentences of this form do not have to be complex in a cognitive sense, the wh-word in the first position makes it more difficult for a DA classifier to deal with correctly. A straightforward non-question on the other hand does not contain features which suggest a question in the first few words.

As with the questions, varying forms of difficult non-questions can be generated from this starting point. Using this type of data calls for some fine judgement. The problem is that it is possible to construct some really taxing non-questions using wh-chefts as pronouns (and find other function word usages that cause problems). However, of their nature some of these are quite unnatural sounding and would be of very low frequency in a natural interaction. So the dilemma arises:

- Should these forms be included in a dataset at all when they are likely to occur with low frequency?
- If so should they be represented in proportion to their frequency of occurrence in the language?

The decisive factor for this work was "Loebner behaviour." This is most prominent in the logs of judging of the annual Loebner prize, although it has been observed to a lesser degree in logs from Conversational Agent tests conducted by the MMU Centre for Conversational Agents. When a human uses a Conversational Agent, the interaction tends to proceed without incident as long as the agent behaves like a human. As conversational partners (particularly in goal-oriented dialogue) both humans and Conversational Agents will make mistakes. As long as the agents make human-like mistakes things go reasonably well. However, even if a Conversational Agent is outperforming the equivalent human in terms of things like domain knowledge, once it makes the kind of mistake a human would not make the user tends to fasten on this to the exclusion of achieving the goal. Judges in the Loebner prize competition tend to be particularly vicious and relentless in this kind of behaviour.

So exploring the issues of "difficult non-questions" is important if the product of the DA classifier is to be deployed ultimately in a Conversational Agent.

4.7.3 Formulation of the general hypothesis for the experiments in this chapter

A large number of experiments were conducted to investigate the performance of DA classifiers for different combinations of the classes of questions and non-questions discussed above. These were intended to address the general research question:

"Are function words an effective source of features for classifying Dialogue Acts?"

For this series of experiments the question has been simplified to:

"Can function words effectively discriminate between certain subsets of questions and non-questions?"

The basic form of the hypotheses for the experiments was:

H_0: Function word features can not discriminate between questions and non-questions
H_1: Function word features can discriminate between questions and non questions

In a more specific form:

H_0: Function word features can not achieve classification accuracy significantly higher than chance over the test dataset of questions and non-questions

H_1: Function word features can achieve classification accuracy significantly higher than chance over the test dataset of questions and non-questions

Specific hypotheses are stated later with their respective experiments.

4.7.4 Collection of raw question / non-question data

Prior experience with the experiments to generate the similarity benchmark dataset suggested that trying to get humans to generate this data would have a low productivity of usable data. Consequently data was collected from web sources. Some investigation revealed that the Usenet newsgroups maintained a list of the highest rated (by the users) Frequently Asked Question (FAQ) lists across the range of groups. These were supplemented by other suitable FAQ lists found using search engines. Table 4.11 describes the FAQ lists used to construct the question set, with the numbers of questions in each list (No.).

Finding suitable non-questions posed certain problems. It was important to avoid introducing spurious features that would artificially boost the discrimination between the questions and non-questions.

FAQ List	Description	No
Ambisonic	Questions about a form of Surround Sound for HiFi enthusiasts	21
BBC Online	Enquiries about programmes, problems with reception etc. posed to the British Broadcasting Corporation	292
BCAB	On using the mailing list of the British Computer Association of the Blind	10
Bipolar	Questions about Bipolar Disorder	30
Chocolate	Various questions about chocolate (cooking etc.)	22
Classical Guitar	Questions about tuition, music resources etc for Classical Guitar	27
CocaCola	Questions (marketing oriented) posed to Coca Cola Great Britain	16
Banks	Questions about mortgages, credit cards, account management etc compiled from multiple Bank websites	271
Supermarkets	Questions about loyalty cards, web shopping, job opportunities etc from multiple supermarket sites	206
Ethics	Questions about ethical issues in organising medical	18

	and psychological experiments	
Gasoline	Technical questions about petrol and internal combustion engines	89
Hair and Beauty	Questions about hair care, wedding and beauty topics from multiple web sources	332
Ikea	Questions about shopping from the IKEA group of furniture stores	92
IRS	Large set of questions posed to the US Internal Revenue Service about tax allowances, obligations and responsibilities	384
JobInterview	A set of questions you are likely to be asked at job interviews	22
NHS Immunisation	Questions about vaccination, who should be vaccinated, travel requirements, side effects etc. from the UK National Health Service	78
Tattoo	Questions about finding tattoo artists, health issues etc.	48

Table 4.11 FAQ question sources

This danger could arise from stylistic differences if the questions and non-questions came from different domains. Therefore one criterion was to derive (as far as possible) non questions using topics related to the questions. The second problem was the nature of the non-questions. As the dataset should mimic dialogue, non-questions were needed that were as close as possible to a conversational style. Looking at web sites in general was unproductive in finding useful material. Despite reservations about the general enthusiasm for mining data from blogs, in this respect they were found to be useful. The best blogs for this purpose are like listening to an internal conversation playing in the author's head. As was the case with the questions, material valued by the relevant community was collected, in this case "blogs of note" on blogspot.com, peer-selected by the blogger team. These were supplemented by other blogs chosen for their overlap with the FAQs. Table 4.12 describes the blogs chosen, with the numbers (No.) of questions extracted from each blog.

Blog	Description	No.
Hifi-blog.com	Discussion, review, answers to questions about HiFi in general	98
www.bbc.co.uk/blogs	Blogs from multiple sources in the British Broadcasting Corporation	352
www.actionforblindpeople.org.uk/text-only/news/blog/	Support / advocacy for blind people	36

http://www.bipolar-lives.com/bipolar-blog.html	News, commentary and support for people with bipolar disorder	81
blog.chocolateadvocate.com	Production and use of chocolate	125
www.classicalguitarblog.net	Promotion / discussion of classical guitar and competitions	136
www.coca-colaconversations.com	By a historian / archivist for Coca Cola	60
Various Pensions Investments blogs	Compiled from a number of bloggers writing on pensions and investments intended to semantically balance Bank FAQs	116
http://foodplacefun.blogspot.com/	Humorous blog about working in a supermarket	436
blogs.guardian.co.uk/science/	Various blogs containing material related to medical ethics	62
Various motoring blogs	Petrol consumption, diesel engines, motoring matters	94
Various hair and beauty blogs	Product promotion and advice	128
ikeahacker.blogspot.com	Advice for Ikeahackers (who customise IKEA furniture)	73
Various financial blogs	These financial blogs have a high content concerned with the IRS	84
http://sethgodin.typepad.com/seths_blog/	Contains a section on job interviews	64
Nhsblogdoc.blogspot.com	About the UK National health Service	88
Various blogs about tattoos	Information about tattoos and issues of tattoo removal	93
http://blackeiffel.blogspot.com	Design blog with an emphasis on weddings / relationships	303
teaforjoy.blogspot.com	Blogs of note design blog	56
Various milblogs	Blogs of note about military issues by serving personnel and veterans	296

Table 4.12 Blog non-question sources

Sampling

Records for the two classes (question, non-question) were created by extracting a subset of the sentences in the class. The selection criterion was to take all of the sentences from the

sources with small numbers (e.g. all 18 Ethics questions, all 36 non-questions from the Blind blog) then randomly sample the sources with large number entries to complete the dataset. Slightly more than 600 were sampled from each class at this stage for the following reason. When a text is tokenised loss of the content word information transforms it into a more generalised form. As a consequence two different source texts may produce the same set of tokens. Therefore the questions and non-questions were tokenised separately and each file was checked for duplicates which were removed. At this stage, if more than 600 records were left in the file excess records were removed at random. Finally the two files were merged.

Samples of the 600 Straightforward Questions, Straightforward Nonquestions and Difficult Non-questions are given in appendix 4.1, appendix 4.2 and appendix 4.3 respectively.

4.7.5 Creating datasets for particular combinations of the question / non-question forms

A strategic decision had to be made which affected the composition of the datasets. The decision was whether to:

- construct a single classifier from data which covered all of the permutations of questions and nonquestions
- construct a number of classifiers, one for each permutation of questions and nonquestions

The first option could require prohibitively large quantities of straightforward questions to be collected if they had to balance both the straightforward and difficult non-questions. Also it was possible that combining the outputs from several independently trained trees (question /nonquestion, instruction/noninstruction etc.) would form the most effective multi-classifier. In any event it was considered that multiple classifiers would provide an acceptable proof of concept.

Proportions of Questions and Nonquestions

There were a number of criteria that the dataset had to satisfy. The most obvious criterion was that of size. The minimum size is determined by the diversity and complexity of the features which have to be represented if the underlying model of the domain is to be represented correctly. Quinlan noted that this could not be determined in advance (Quinlan, 1993), but mentioned that it could lie between hundreds and thousands of training cases.

Based on this and experience of other unsupervised learning applications, it was decided to start with an evenly split dataset of 600 questions and 600 non questions. Even though this may not reflect the frequency of occurrence of the classes in the real world, it is a commonly accepted practice which prevents bias in training and testing. Using the n-fold cross validation technique (described later) this dataset would allow in excess of 1,000 training cases to be used.

Following some pilot experimental work, it was decided to use data in equal proportions and to perform separate evaluations for straightforward questions vs. straightforward non-questions and straightforward questions vs. difficult non-questions.

Therefore, to perform the planned investigation two sets of 600 non-questions were required – straightforward non-questions and difficult non-questions. Furthermore, the questions had to be straightforward but not trivial, which ruled out adopting TREC questions for the dataset. It might have proved possible to mine suitable questions from one of the DA collections – but this proposition was not convincing because examples in the reviewed papers looked semantically quite trivial with the main challenges coming as a result speech recognition errors, fragmentary utterances or the distribution of the data over multiple dialogue channels. However the most promising potential datasets tended to have very high subscription fees. Consequently a new dataset was collected for the experiments in this work.

4.7.6 Preprocessing the datasets for feature extraction

A successful DT relies on an appropriately designed feature set. Prior work on DA classification has used some sophisticated features but these required manual extraction and could not be automated.

This work uses simple features. Each word position in the sentence is coded as a specific function word, or if no function word is found as a generic content word.

This type of feature is very simple for a DT to process, provided certain requirements are considered during feature extraction. These are:

- Preprocessing to remove punctuation and spacing
- Expanding contractions
- Extending to a fixed length record
- Feature extraction using an appropriate numerical scale

4.7.6.1 Removing punctuation and spacing

The first stage of pre-processing is cleaning the data by removing punctuation; additional spaces etc. to produce a string of words separated by single spaces. Removing punctuation has a side effect on the treatment of compound terms. Examples include the website greatamericans.com being converted to greatamericanscom and, apparently more seriously, behind-the-scenes being converted to behindthescenes. However, on examination this does not appear to pose a problem for the technique. Investigation of the compound terms reveals that they are fundamentally "super-content words". For example greatamericans.com is the name of a website and therefore a noun. In the phrase "I am just a lowly behind-the-scenes guy" the term is an adjective. In all cases the removal of the hyphens results in a word which is not recognisable as a function word and is therefore, correctly classified as a content word. The occurrence of "the" in behind-the-scenes is not as a function word per se, but as glue, holding the super-adjective together.
There is one exception to the removal of punctuation marks. The apostrophe in possessives is allowed through. Apostrophes are given special treatment because they are used for expanding contractions during pre-processing. However, the semantic information in the possessives is not used and it is discarded at the tokenisation stage. This is one illustration of the quite brutal approach taken in pre-processing where efficiency is prioritised over elegance.

4.7.6.2 Expanding contractions

Contractions are forms such as *can't*. These are important because they are composed of function words - the basis of the technique used here. Expansion of *can't* replaces it with the words *can not*. This is performed using a lookup table (shown in appendix 4.4), which again is efficient to implement. Some contractions cannot be expanded deterministically using this technique; this is the apostrophe-s problem. Consider the following for example:

> Whatever's the matter?
> Whatever's been going on?

Human intuition tells that in the first case the expansion is *whatever is* and in the second it should be *whatever has*. This is a disambiguation problem, traditionally tackled with NLP techniques starting with parsing (which could also introduce errors of its own). As a major driving force of this work is the production of computationally light-weight algorithms alternative approaches must be sought. A second option would be to train another DT to

perform the disambiguation. However appealing this may be, it must be rejected as potentially introducing a confounding factor into the following experiments. Therefore the tree will receive some incorrect expansions but may learn to cope with them – this will be determined by the experiments which follow.

4.7.6.3 Extending to a fixed length record

Deciding on the features is one of the crucial design aspects of AI-based classifiers. The issue is to retain the information contained in the original data from which the classification is made, whilst helping the classifier to determine which features contribute to classification and incorporate them in the model. Some decisions are forced upon us. For example, DTs (and therefore the algorithms that implement them) generally work with fixed length records. Therefore one must choose a maximum record size even though the utterances can be of variable length up to that maximum.

Spoken dialogue utterances can be very short – in many cases the utterances are less than 5 words (van Noord, 1997, He and Young, 2005). Typed messaging-based systems tend to produce longer utterances, not having the immediacy of spoken dialogue. In creating the STSS similarity test data limits of 10-20 words were set for the data items. Bearing this in mind, it was decided to set the upper limit of length to be 25 words in this series. This allows for a 20 word propositional content plus a generous allowance of 5 words for extra structural words required for the question.

There is a point to be made very clearly here. The 25 word length criterion has been set for this particular experimental series. In collecting the data a number of significantly longer text-based question and non-question items were found. However, the majority of items collected were within this upper bound and the vast majority of speech-based dialogue utterances are within the limit. More importantly, the experiments that follow do not stand or fall on the choice of this particular parameter. If longer utterances must be catered for it is simply a matter of training a DT with the appropriate data and parameters. One factor which is important is that the technique must cover all utterance lengths up to the maximum size. The consequence of the above is that there are 25 features, corresponding to the word positions in the utterance (labelled W1 to W25).

4.7.6.4 Feature extraction

Feature extraction takes place by analysing the pre-processed string one word at a time and replacing the word by a numerical token at the output. In the algorithm, a 2-D array of

function words and numeric tokens is searched for the word. Table 4.13 shows extracts from this array, which has 264 rows.

A	1
about	2
above	3
. . .	. . .
myself	131
namely	132
. . .	. . .
Yours	262
yourself	263
yourselves	264

Table 4.13 A sample of function words and tokens extracted from the data structure

The full listing of this table is given in appendix 4.5. If the word is found, the appropriate token (range 1 – 264) is inserted in the output string. In this experiment the tokens are allocated in ascending order to the alphabetically sorted list of words, so that 1 represents the word "a" and 264 represents the word "yourselves". There may be some underlying structure that can be exploited by using a different ordering of the function words and this will be investigated later. However, the prudent approach in AI is to make no presumptions at the start of a series of experiments therefore alphabetical order was chosen for the experiments in this chapter.

Any word which is not in the function word list is assumed to be a content word. Every content word is replaced by the token 0. This maintains the correct spatial relationships between the function words in the utterance. Finally, when an utterance contains less than 25 words, the remaining word positions in the token are filled with the "no word present" token, 300. An example of a tokenised question is given in table 4.14.

Question	does wearing caps or hats contribute to hair loss
Tokenised Form	56,0,0,156,0,0,212,0,0,300,300,300,300,300,300,300,300,300,300,300, 300,300,300,300,300

Table 4.14 Tokenisation of training data

So in the tokenised form, the function word "does" is replaced by token 56 from the function word table. The content words "wearing caps" are not founding in the table so each is replaced by 0, the function word "or" is replaced by token 156 from the table etc. When the final word "loss" has been replaced by 0, there are fewer than 25 tokens so the rest are filled with the token 300.

Why assign these particular values to content words and empty slots? In the C4.5 algorithm, an empty slot will be treated as a missing attribute. Consequently it propagates weights down multiple branches of the tree to cope with the problem. This is actually a subtle misinterpretation of the data. In fact the empty slots provide information about the length of the utterance therefore the "empty slot" token with a value of 300 was created. The token 300 was chosen to allow room for the function word array to grow (in case some function words had been overlooked) and still leave room to split between the function word tokens and the no word present token. The value of 0 for content words was chosen for two reasons. First, putting it at one extreme of the range allows an easy split between content words and everything else (all the function words and the "no word present" token). Second, convenience – numbering from 1 to n was suitable for array indexing leaving zero free for that particular token.

4.7.6.5 Token scale type

It is important to emphasise that the coding using numeric tokens does not constitute a mathematically significant measurement scale. The formal designation of the scale type is nominal, but using numeric labels. The items are **not** on a monotonically increasing scale corresponding to some quantifiable semantic or syntactic property – only alphabetical order. They are merely 266 discrete items which have been given numeric rather than text labels. So when the DT makes splits like:

w5 <= 21

it is merely inferring that there is some value to splitting the words into two groups – one from *a* to *anyhow* and the other from *anyone* to *yourselves*.

The only other consideration is that the content word and no word present tokens are at extreme ends of the range. If the first level of splitting would work better by keeping them together, this advantage will be missed and it will be an interesting topic for future work. To summarise, the coding scheme described above has been chosen because it is easy to define objectively, efficient to apply and simple to implement.

4.8 Dataset creation

The specific datasets for each of the experiments that follow were derived from the sets of 600 Straightforward Questions, 600 Straightforward Non-questions and 600 Difficult Nonquestions described in section 4.7.2 (samples provided in appendix 4.1, 4.2, 4.3 respectively). Different combinations of these were used, e.g. straightforward questions / straightforward nonquestions, straightforward questions / difficult nonquestions etc.

4.8.1 Preambles

In a number of experiments the texts are pre-fixed with a preamble. The preambles were generated by a retired English teacher according to supplied instructions and there were 50 preambles for each possible length. To avoid artefacts introduced by ordering, in every case the original dataset was randomised before pre-fixing the preambles, which were then applied in repeated blocks. The preambles are shown in appendices 4.6 (1-word), 4.7 (2-word) and 4.8 (3-word).

One dataset used a combination of 1, 2 and 3 word preambles. This resulted in 400 items (200 questions / 200 non-questions) with each length of preamble, towards the lower limit of training data requirements discussed by Quinlan (Quinlan, 1993). Another series of experiments was run using a single length of preamble for lengths from 1 to 3 words. This posed another problem. As there were 50 preambles for each possible length, each preamble appeared 12 times in any dataset it was applied to. This raised the research question "Will the DT classifier learn spurious features from the preambles?" This could be investigated in two ways: by examining the structure of the generated DTs and by running additional experiments in which both questions and non-questions had preambles attached.

There was an interesting possibility of the noise factor introduced by the preambles combining with the distraction factor of the difficult non-questions to further affect Classification Accuracy. A series of runs was conducted using each of the three forms of pre-amble from the previous experiments applied to both the straightforward questions / difficult non-questions combination.

4.8.2 Simulated clauses

It was decided to investigate some of the challenges involved in dealing with tag questions and questions composed of multiple sentences. However, the collection and classification

of training and testing datasets will pose a serious research challenge in its own right. Therefore it was decided to use a simulated dataset in this chapter. This dataset was called "simulated clauses" and consisted of the 600 straightforward questions and the 600 straightforward nonquestions prefixed by difficult non-questions. In each case a different difficult non-question was used as the prefix. Even though the resulting combinations were somewhat unusual, they provided a realistic challenge to investigate the effectiveness of the technique in dealing with the tag and multiple sentences forms of question.

4.9 Specific hypotheses

The particular experiments, described in table 4.15, are grouped into series of 4 each of which addresses a particular variant of the general hypothesis. Each group of 4 experiments, attached to a particular hypothesis, uses its own dataset as described above.

Experiments	Hypothesis (H_1)
4.1-4.4	A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions.
4.5-4.8	A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions with 1 -3 word preambles vs. straightforward non-questions.
4.9-4.12	A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions with 1 word preambles vs. straightforward non-questions.
4.13-4.16	A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions with 2 word preambles vs. straightforward non-questions.
4.17-4.20	A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions with 3 word preambles vs. straightforward non-questions.
4.21-4.24	A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions when both have 1 word preambles. (The decision tree does not learn to classify more effectively by learning from features in 1-word preambles)
4.25-4.28	A decision tree using function words can achieve classification

	accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions when both have 2 word preambles. (The decision tree does not learn to classify more effectively by learning from features in 2-word preambles)
4.29-4.32	A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions when both have 3 word preambles. (The decision tree does not learn to classify more effectively by learning from features in 3-word preambles)
4.33-4.36	A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. difficult non-questions.
4.37-4.40	A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. difficult non-questions when both have 1 word preambles.
4.41-4.44	A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. difficult non-questions when both have 2 word preambles.
4.45-4.48	A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. difficult non-questions when both have 3 word preambles.
4.49-4.52	A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions when both are preceded by difficult simulated clauses.

Table 4.15 experimental series

4.10 Experimental parameters

4.10.1 Baseline measures and statistical significance

The form of the data will always be 600 questions / 600 non-questions in the the experiments in this chapter, so the chance level of Classification Accuracy is 50%.
In the following experiments, each run (e.g. experiment 4.1) produces a series of DTs using different parameters. A feature of Weka is that it uses a particular set of parameter values as a baseline and reports the statistical significance of variations from this CA. This is performed using the corrected re-sampled *t*-test.

The important point about this is that all of the experiments were arranged so that baseline trees used for comparison were created using confidence = 0.25 and minimum number of objects = 2, even if these trees do not appear in a table of results.
In the tables of results a significant decrease in CA is marked with a * and a significant increase with the letter v.

4.10.2. N-fold Cross Validation

All of the experiments that follow make use of n-fold cross validation. N-fold cross validation is a technique which optimises the use of the dataset for training. It allows every item in the dataset to participate as both training and test data whilst ensuring that no item can be used for training and testing on the same classifier
This is achieved by allocating the data to n folds at random, so for 10-fold cross validation the data would be randomly partitioned into 10 equally sized folds (allowing for minor variation where the dataset is not an integral multiple of the number of folds). Using an ordinary training and test set split, a single classifier would be constructed for each set of experimental conditions (although this could be repeated many times with randomised starting conditions, averaging the results to obtain more robust results).
Under 10-fold cross validation that single classifier is replaced by 10 classifiers. A different fold is used as the test data for each of these classifiers. But when that particular fold is used for testing the other 9 are used for training (fold 1 vs. folds 2 – 10, fold 2 vs. folds 1 and 3 - 10 etc.). So a result reported for 10-fold cross validation is in fact a composite of the information from 10 runs. 10-fold cross validation is used frequently and has become a convention, probably because it is the default setting in many packages. However, choice of the number of folds is up to the designer of the experiment, who has a responsibility to choose *n* based on estimates of the complexity of the problem and the size of the available data. The extreme upper limit on folds is sometimes used in psychology experiments where collection of data is difficult and the domain is not well understood. This is known as "leave one out re-sampling" and involves using each individual item as a test case against classifiers trained with all of the remaining data. This technique has been used with datasets containing less than 100 items (Rothwell et al., 2006).
60-fold cross validation was used for the experiments in this work. This was chosen because it allows the use of 1,180 cases for training and 20 for testing in each variant. Over 1000 training cases is in line with the observation about datasets for training C4.5 made by Quinlan (Quinlan, 1993) and the computational requirements are reasonable convenient for a current dual-core cpu computer. Observation of the distribution of the features in the data

suggests that the distribution may be quite sparse and complex, supporting the decision to use larger number of training cases.

Using the Weka experimenter, runs can be set up which repeat the whole series of the experiments in the run 10 times and present results consolidated for the whole set also within a reasonable time. Consequently each result reported in the results tables that follow represents the construction and testing of 600 DTs. This is valuable as randomisation of starting conditions means an exhaustive exploration of the domain takes place.

4.10.3 Pruning

All the discussion so far has concerned the construction of a classifier which can achieve good classification accuracy. However, there are two other issues which are addressed by pruning the tree. The most common concern is the issue of modelling the domain as opposed to just modelling the dataset; accordingly one concern of this work is how compact the tree can be made before its Classification Accuracy drops significantly.

The second issue is one of computational efficiency. Pruned trees require fewer decisions to be taken to reach a classification resulting in greater efficiency. This work is unusual because the anticipated applications potentially involve very large numbers of concurrent users connected via the internet. Also, DA classification may involve multiple classifiers co-operating to classify the DAs. Therefore, optimal pruning is particularly important.

For each hypothesis to be tested the results of 4 different runs will be reported. Two runs will be for Confidence Level pruning and two for Minimum Number of Objects Pruning. The first run is intended to find roughly where in the range Classification Accuracy changes significantly from the established baseline value. The second is intended to pin this point down more precisely. Occasionally the point where Classification Accuracy has changed significantly is not found by the first run; in those cases intermediate runs (results not reported) are used to establish a suitable point for the second run.

The standard parameters which will be used for Confidence Levels are 0.25, 0.2, 0.15, 0.1, 0.05 and the parameters for MNO are 2, 5, 10, 15, and 20. For all of the runs using Confidence Level pruning, the baseline is established by the result of the 0.25 value. For all of the runs using MNO, the baseline is established by the 2 value.

The tree size parameter is the total number of nodes in the tree, counting a leaf as a node.

In the tables that follow the abbreviation **CA** is used to represent the term Classification Accuracy (not Conversational Agent).

4.11 Results of Experiments

The following sets of experiments are presented in groups of 4. Each group of 4 corresponds to one combination of question / non question data listed in table 4.15. The experiments show how the maximum classification accuracy and the best pruned decision tree size obtainable for each combination was determined, with the function words sorted into alphabetical order. Two of the 4 tables from each group show the outcome of confidence interval pruning and two show the results of minimum number of objects pruning.

In the following tables Conf refers to the confidence interval, Min refers to the minimum number of objects allowed in a leaf, CA refers to the percentage classification accuracy and Tree size refers to the number of nodes in a tree.

Experiments 4.1 to 4.4 (straightforward questions / straightforward non-questions, no preambles)

H_0: A decision tree using function words can not achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions.

H_1: A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	98.50	98.51	98.41	98.48	98.48
Tree Size	29-71	29-67	29-47	29-33	29-33

Table 4.16 Experiment 4.1

Conf	0.0005	0.0004	0.0003	0.0002	0.0001
CA	98.36	98.36	97.88*	97.89	98.77
Tree Size	25-31	25-31	21-31	21-31	17-31

Table 4.17 Experiment 4.2

Min	2	5	10*	15	20
CA	98.50	98.32	97.16	97.03	93.57
Tree Size	29-71	25-47	21-25	19-29	15-25

Table 4.18 Experiment 4.3

Min	5	6	7*	8	9
CA	98.32	98.28	97.80	97.62	97.21
Tree Size	25-47	25-49	25-35	21-35	21-35

Table 4.19 Experiment 4.4

Findings

The highest CA achieved was 98.51. The tree is performing significantly better than chance and by a very large margin. This provides good evidence to reject the null hypothesis and accept that the tree is a good classifier.

The baseline range of tree sizes (Conf = 0.25, MNO = 2) was 29-71 so the smallest were quite compact. Pruning achieved a modest improvement in the lower limit (25) and a good reduction in the upper limit (31) before a significant reduction in CA.

Discussion

This suggests that the tree was already generalising well and if speed is an issue there is a small gain from pruning.

Experiments 4.5 – 4.8 (straightforward questions with a mix 1 – 3 word preambles / straightforward non-questions)

H_0: A decision tree using function words can not achieve classification accuracy significantly higher than chance over the dataset of straightforward questions with 1 – 3 word preambles vs. straightforward non-questions.

H_1: A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions with 1 -3 word preambles vs. straightforward non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	79.17	79.22	79.11	78.22	78.54
Tree Size	131-211	113-189	103-157	87-141	59-123

Table 4.20 Experiment 4.5

Conf	0.000005	0.000004	0.000003	0.000002	0.000001
CA	78.83	78.47	77.89	76.53*	73.79
Tree Size	15-35	15-35	13-33	9-31	9-31

Table 4.21 Experiment 4.6

Min	2	5	10	15	20
CA	79.17	78.46	78.20	77.02	76.16*
Tree Size	131-211	69-117	39-71	29-57	27-43

Table 4.22 Experiment 4.7

Min	15	16	17	18	19	20
CA	77.02	76.95	76.76*	76.33	76.43	76.16
Tree Size	29-57	29-53	25-49	25-47	27-43	27-43

Table 4.23 Experiment 4.8

Findings

The highest CA reported was 79.22**.** The tree is performing significantly better than chance by a large margin. There is good evidence to reject the null hypothesis and accept that the tree is a good classifier.

The baseline range of tree sizes (Conf = 0.25, MNO = 2) was 131-211. Pruning achieved a substantial improvement with a range from 13-33 before a significant reduction in CA.

Discussion

This suggests that pruned trees should generalise well and that suitable speed of execution can be achieved.

The first word in a straightforward question provides information that it is a question. The purpose of the preambles was to increase the challenge to the classifier by adding noisy information (a mix of function and content words) in front of that word.

Even though the classifier performed well, the drop in CA reflects the fact that the dataset has increased in complexity. The following 3 experiments investigate the use of a separate classifier for each of the 3 lengths of preamble, which could be used in a multi-classifier system.

Experiments 4.9-4.12 (straightforward questions with 1 word preambles / straightforward non-questions)

H_0: A decision tree using function words can not achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions with a 1 word preamble.

H_1: A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions with 1 word preambles vs. straightforward non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	87.12	87.67	87.86	88.00	88.11
Tree Size	69-133	59-115	47-103	43-95	31-91

Table 4.24 Experiment 4.9

Conf	0.0005	0.0004	0.0003	0.0002	0.0001
CA	86.13	85.63	85.33	84.45*	84.27
Tree Size	11-39	11-39	11-39	9-39	9-29

Table 4.25 Experiment 4.10

Min	2	5	10	15	20
CA	87.12	87.00	87.08	85.95	85.25*
Tree Size	69-133	45-83	23-55	23-43	17-33

Table 4.26 Experiment 4.11

Min	15	16	17	18	19	20
CA	85.95	85.66	85.75	85.46	84.87*	85.25
Tree Size	23-43	23-43	19-37	19-35	17-33	17-33

Table 4.27 Experiment 4.12

Findings

The highest CA achieved was 88.11**.** The tree is performing significantly better than chance and by a large margin. This provides good evidence to reject the null hypothesis and accept that the tree is a good classifier.

The baseline range of tree sizes (Conf = 0.25, MNO = 2) was 69-133. After pruning there was a substantial improvement with a range from 11-39 before a significant reduction in CA.

Discussion

This suggests that pruned trees should generalise well and that suitable speed of execution can be achieved.

The CA of 88.11 is substantially higher than the previous classifier's value of 79.22. This evidence supports the position that a multiple classifier would be effective for mixed-length preambles.

Experiments 4.13-4.16 (straightforward questions with 2 word preambles / straightforward non-questions)

H_0: A decision tree using function words can not achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions with a 2 word preamble.

H_1: A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions with 2 word preambles vs. straightforward non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	89.40	89.55	89.62	88.98	89.23
Tree Size	73-135	71-125	53-111	43-101	33-87

Table 4.28 Experiment 4.13

Conf	0.00005	0.00004	0.00003	0.00002	0.00001
CA	88.01	87.93	87.88	87.10*	85.76
Tree Size	9-13	9-13	9-13	9-13	9-13

Table 4.29 Experiment 4.14

Min	2	5	10	15	20
CA	89.40	89.00	87.83	87.92	87.91*
Tree Size	73-135	49-89	25-59	21-43	17-33

Table 4.30 Experiment 4.15

Min	25	26	27	28	29	30
CA	88.19	87.85	87.12*	87.11	86.91	86.68
Tree Size	15-31	11-31	11-29	11-25	11-25	11-25

Table 4.31 Experiment 4.16

Findings

The highest CA achieved was 89.62**.** The tree is performing significantly better than chance by a large margin. This provides good evidence to reject the null hypothesis and accept that the tree is a good classifier.

The baseline range of tree sizes (Conf = 0.25, MNO = 2) was 73-135. After pruning there was a substantial improvement with a range from 9-13 before a significant reduction in CA.

Discussion

This suggests that pruned trees should generalise well and that suitable speed of execution can be achieved.

The CA shows a slight improvement in performance on the classifier for questions with 1-word preambles and again is substantially higher than the 79.22 achieved by the classifier

for mixed length preambles. This evidence supports the position that a multiple classifier would be effective for mixed-length preambles.

Experiments 4.17-4.20 (straightforward questions with 3-word preambles / straightforward non-questions)

H_0: A decision tree using function words can not achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions with a 3 word preamble.
H_1: A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions with 3 word preambles vs. straightforward non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	88.86	88.72	88.64	88.73	88.75
Tree Size	71-131	65-113	57-103	53-97	37-89

Table 4.32 Experiment 4.17

Conf	0.007	0.006	0.005*	0.004	0.003
CA	87.77	87.44	86.75	86.64	86.76
Tree Size	15-49	15-41	15-35	15-35	11-33

Table 4.33 Experiment 4.18

Min	2	5	10	15	20
CA	88.86	87.88	88.37	87.59	88.23
Tree Size	71-131	45-81	31-49	23-39	23-33

Table 4.34 Experiment 4.19

Min	45	46	47	48	49
CA	87.84	87.04	86.58*	86.63	86.60
Tree Size	11-13	11-13	11-13	11-13	11-13

Table 4.35 Experiment 4.20

Findings

The highest CA achieved was 88.86. The tree is performing significantly better than chance by a large margin. There is good evidence to reject the null hypothesis and accept that the tree is a good classifier.

The baseline range of tree sizes (Conf = 0.25, MNO = 2) was 71-131. After pruning there was a substantial improvement with a range from 9-13 before a significant reduction in CA.

Discussion

This suggests that pruned trees should generalise well and that suitable speed of execution can be achieved.

The CA falls between that for the 1-word preamble classifier and that for the 2-word preamble classifier. It is substantially higher than the 79.22 achieved by the DT for mixed length preambles. Again this evidence supports the position that a multiple classifier would be effective for mixed-length preambles.

Experiments 4.21-4.24 (straightforward questions / straightforward non-questions both with 1 word preambles)

H_0: A decision tree using function words can not achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions when both have1 word preambles.

H_1: A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions when both have 1 word preambles.

Conf	0.25	0.2	0.15	0.1	0.05
CA	98.24	98.38	98.39	98.39	98.48
Tree Size	31-45	25-41	25-33	25-33	25-33

Table 4.36 Experiment 4.21

Conf	0.00006	0.00005	0.00004	0.00003	0.00002
CA	97.70	97.66	97.48*	97.40	97.40
Tree Size	13-31	13-31	13-31	13-31	13-31

Table 4.37 Experiment 4.22

Min	2	5	10	15	20
CA	98.24	98.53	97.39*	97.27	97.08
Tree Size	31-45	27-29	21	17-21	17-21

Table 4.38 Experiment 4.23

Min	5	6	7	8	9
CA	98.53	98.48	97.96	97.38*	97.39
Tree Size	27-29	25-29	21-25	21-27	21-27

Table 4.39 Experiment 4.24

Findings

The highest CA achieved was 98.53**.** The tree is performing significantly better than chance and by a very large margin. This provides good evidence to reject the null hypothesis and accept that the tree is a good classifier.

The baseline range of tree sizes (Conf = 0.25, MNO = 2) was 31-45, so the trees were reasonably compact. Pruning achieved an improvement in the lower limit (21) and a good reduction in the upper limit (25) before a significant reduction in CA.

Discussion

This suggests that pruned trees should generalise well and that suitable speed of execution can be achieved.

In this experiment the CA is very close to the value for the straightforward questions and non-questions, providing evidence that the DT has learned to ignore the preambles.

Experiments 4.25-4.28 (straightforward questions / straightforward non-questions both with 2 word preambles)

H_0: A decision tree using function words can not achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions when both have 2 word preambles.

H_1: A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions when both have 2 word preambles.

Conf	0.25	0.2	0.15	0.1	0.05
CA	98.51	98.53	98.48	98.51	98.51
Tree Size	29-73	29-73	29-47	29-33	29-33

Table 4.40 Experiment 4.25

Conf	0.004	0.003	0.002	0.001	0.0009
CA	98.43	98.43	98.43	97.88*	97.88
Tree Size	21-33	21-33	21-33	21-31	21-31

Table 4.41 Experiment 4.26

Min	2	5	10	15	20
CA	98.51	98.33	97.27*	97.16	93.63
Tree Size	29-73	25-55	21-29	17-23	15-23

Table 4.42 Experiment 4.27

Min	5	6	7	8	9
CA	98.33	98.19	97.67*	97.26	97.18
Tree Size	25-55	25-49	21-39	21-37	21-31

Table 4.43 Experiment 4.28

Findings

The highest CA reported was 98.53**.** The tree is performing significantly better than chance and by a very large margin. There is good evidence to reject the null hypothesis and accept that the tree is a good classifier.

The baseline range of tree sizes (Conf = 0.25, MNO = 2) was 29-73. After pruning there was a modest improvement in the lower limit (21) and a good reduction in the upper limit (33) before a significant reduction in CA.

Discussion

This suggests that the tree was already generalising to model the data well and if speed is an issue there is a small gain from pruning.

In this experiment the CA is very close to the value for the straightforward questions and non-questions, providing evidence that the DT has learned to ignore the preambles.

Experiments 4.29-4.32 (straightforward questions / straightforward non-questions both with 3 word preambles)

H_0: A decision tree using function words can not achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions when both have 3 word preambles.

H_1: A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions when both have 3 word preambles.

Conf	0.25	0.2	0.15	0.1	0.05
CA	98.42	98.43	98.30	98.44	98.40
Tree Size	29-65	29-61	29-61	29-33	25-37

Table 4.44 Experiment 4.29

Conf	0.003	0.002	0.001	0.0009	0.0008
CA	98.35	98.35	97.73*	97.73	97.73
Tree Size	25-33	25-33	21-33	21-33	21-33

Table 4.45 Experiment 4.30

Min	2	5	10	15	20
CA	98.42	97.90	97.30*	96.99	92.46
Tree Size	29-65	25-55	21-29	17-25	15-23

Table 4.46 Experiment 4.31

Min	5	6	7	8	9
CA	97.90	97.72	97.29*	96.81	96.87
Tree Size	25-55	25-53	25-45	21-43	21-31

Table 4.47 Experiment 4.32

Findings

The highest CA achieved was 98.44**.** The tree is performing significantly better than chance by a very large margin. This provides good evidence to reject the null hypothesis and accept that the tree is a good classifier.

The baseline range of tree sizes (Conf = 0.25, MNO = 2) was 29-65. Pruning achieved a modest improvement in the lower limit (25) and a good reduction in the upper limit (33) before a significant reduction in CA.

Discussion

This suggests that the tree was already generalising to model the data well and if speed is an issue there is a small gain from pruning.

In this experiment the CA is very close to the value for the straightforward questions and non-questions, providing evidence that the DT has learned to ignore the preambles.

Experiments 4.33-4.36 (straightforward questions / difficult non-questions no preambles)

H_0: A decision tree using function words can not achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. difficult non-questions.

H_1: A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. difficult non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	89.18	89.55	89.50	89.36	88.93
Tree Size	75-139	49-123	45-105	45-101	41-89

Table 4.48 Experiment 4.33

Conf	0.0004	0.0003	0.0002	0.0001	0.00009
CA	87.93	87.94	87.77	87.38 *	87.34
Tree Size	25-39	25-39	23-37	23-37	23-37

Table 4.49 Experiment 4.34

Min	2	5	10	15	20
CA	89.18	89.00	88.41	87.95	87.41 *
Tree Size	75-139	47-85	29-53	25-39	23-33

Table 4.50 Experiment 4.35

Min	14	15	16	17	18
CA	87.83	87.95	87.44 *	87.60	87.71
Tree Size	27-41	25-37	25-35	25-33	25-33

Table 4.51 Experiment 4.36

Findings

The highest CA achieved was 89.55 (although CA was increasing in experiment 4.35 further investigation found it did not reach 89.55). The tree is performing significantly better than chance by a large margin. This provides good evidence to reject the null hypothesis and accept that the tree is a good classifier.

The baseline range of tree sizes (Conf = 0.25, MNO = 2) was 75-139 After pruning there was a substantial improvement with a range from 23-37 before a significant reduction in CA.

Discussion

This suggests that pruned trees should generalise well and that suitable speed of execution can be achieved.

The difficult non-question has a first word which is usually an indication of a question. This has reduced the CA by 8.96 compared with the straightforward questions /

straightforward non-questions. However the best CA of 89.55 for this DT provides clear evidence that it has discovered more complex features to perform the classification.

Experiments 4.37-4.40 (straightforward questions / difficult non-questions both with 1 word preambles)

H_0: A decision tree using function words can not achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. difficult non-questions when both have 1 word preambles.
H_1: A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. difficult non-questions when both have 1 word preambles.

Conf	0.25	0.2	0.15	0.1	0.05
CA	89.13	89.57	89.61	89.93	89.23
Tree Size	77-127	55-117	49-115	45-95	41-89

Table 4.52 Experiment 4.37

Conf	0.00005	0.00004	0.00003	0.00002	0.00001
CA	87.54	87.54	87.54	87.43	86.13 *
Tree Size	23-35	23-35	23-35	23-35	21-35

Table 4.53 Experiment 4.38

Min	2	5	10	15	20
CA	89.13	88.91	88.59	88.21	87.85
Tree Size	7	47-79	33-55	25-39	23-31

Table 4.54 Experiment 4.39

Min	21	22	23	24	25
CA	87.94	87.98	86.88*	86.46	86.39
Tree Size	23-31	23-31	23-31	23-29	23-29

Table 4.55 Experiment 4.40

Findings

The highest CA achieved was 89.93**.** The tree is performing significantly better than chance by a large margin. This provides good evidence to reject the null hypothesis and accept that the tree is a good classifier.

The baseline range of tree sizes (Conf = 0.25, MNO = 2) was 77-127. After pruning there was a substantial improvement with a range from 23-31 before a significant reduction in CA.

Discussion

This suggests that pruned trees should generalise well and that suitable speed of execution can be achieved.

The difficult non-question has a first word which is usually an indication of a question. This has reduced the CA by 8.55 compared with the straightforward questions / straightforward non-questions. However the best CA of 89.93 for this DT provides clear evidence that it has discovered more complex features to perform the classification.

Experiments 4.41-4.44 (straightforward questions/difficult non-questions both with 2 word preambles)

H_0: A decision tree using function words can not achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. difficult non-questions when both have 2 word preambles.

H_1: A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. difficult non-questions when both have 2 word preambles.

Conf	0.25	0.2	0.15	0.1	0.05
CA	88.28	88.90	89.02	89.03	88.69
Tree Size	77-145	59-123	53-119	53-107	41-85

Table 4.56 Experiment 4.41

Conf	0.00005	0.00004	0.00003	0.00002	0.00001
CA	87.08	87.08	87.08	87.04	85.95 *
Tree Size	23-43	23-43	23-43	23-43	21-43

Table 4.57 Experiment 4.42

Min	2	5	10	15	20
CA	88.28	88.96	88.33	87.68	87.46
Tree Size	77-145	51-93	33-55	25-39	23-31

Table 4.58 Experiment 4.43

Min	21	22	23	24	25
CA	87.46	87.68	87.59	86.80	86.22 *
Tree Size	23-31	23-31	23-29	21-29	23-29

Table 4.59 Experiment 4.44

Findings

The highest CA achieved was 89.03**.** The tree is performing significantly better than chance by a large margin. This provides good evidence to reject the null hypothesis and accept that the tree is a good classifier.

The baseline range of tree sizes (Conf = 0.25, MNO = 2) was 77-145 After pruning there was a substantial improvement with a range from 21-29 before a significant reduction in CA.

Discussion

This suggests that pruned trees should generalise well and that suitable speed of execution can be achieved.

The difficult non-question has a first word which is usually an indication of a question. This has reduced the CA by 9.5 compared with the straightforward questions / straightforward non-questions. However the best CA of 89.03 for this DT provides clear evidence that it has discovered more complex features to perform the classification.

Experiments 4.45-4.48 (straightforward questions/difficult non-questions both with 3 word preambles)

H_0: A decision tree using function words can not achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. difficult non-questions when both have 3 word preambles.

H_1: A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. difficult non-questions when both have 3 word preambles.

Conf	0.25	0.2	0.15	0.1	0.05
CA	88.95	89.33	89.04	89.28	88.49
Tree Size	77-135	53-121	51-113	47-97	39-95

Table 4.60 Experiment 4.45

Conf	0.0004	0.0003	0.0002	0.0001	0.00009
CA	87.84	87.85	87.65	87.23 *	87.16 *
Tree Size	25-43	25-43	23-43	23-39	23-39

Table 4.61 Experiment 4.46

Min	2	5	10	15	20
CA	88.95	88.23	88.65	87.78	87.63
Tree Size	77-135	47-81	31-55	23-39	23-31

Table 4.62 Experiment 4.47

Min	21	22	23	24	25
CA	87.93	87.92	86.95 *	86.47 *	86.37 *
Tree Size	23-31	25-35	23-29	23-29	23-29

Table 4.63 Experiment 4.48

Findings

The highest CA achieved was 89.33**.** The tree is performing significantly better than chance by a large margin. This provides good evidence to reject the null hypothesis and accept that the tree is a good classifier.

The baseline range of tree sizes (Conf = 0.25, MNO = 2) was 77-135. After pruning there was a substantial improvement with a range from 23-43 (Confidence) or 25-35 (MNO) before a significant reduction in CA.

Discussion

This suggests that pruned trees should generalise well and that suitable speed of execution can be achieved.

The difficult non-question has a first word which is usually an indication of a question. This has reduced the CA by 9.09 compared with the straightforward questions /

straightforward non-questions. However the best CA of 89.33 for this DT provides clear evidence that it has discovered more complex features to perform the classification.

Experiments 4.49-4.52 Simulated Clauses (difficult non-questions inserted in front of straightforward questions / straightforward non-questions)

H_0: A decision tree using function words can not achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions when both are preceded by difficult simulated clauses.
H_1: A decision tree using function words can achieve classification accuracy significantly higher than chance over the dataset of straightforward questions vs. straightforward non-questions when both are preceded by difficult simulated clauses.

Conf	0.25	0.2	0.15	0.1	0.05
CA	62.13	62.25	62.28	62.91	63.80
Tree Size	81-313	63-301	61-281	51-255	3-173

Table 4.64 Experiment 4.49

Conf	1.0E-5	7.5E-6	5.0E-6	2.5E-6	1.0E-6
CA	66.43 v	66.47 v	66.48 v	66.62 v	66.62 v
Tree Size	3-11	3-11	3-11	3-11	3-11

Table 4.65 Experiment 4.50

Min	2	5	10	15	20
CA	62.13	63.38	63.48	63.91	65.09 v
Tree Size	81-313	37-183	29-101	19-75	15-55

Table 4.66 Experiment 4.51

Min	27	28	29	30	31
CA	63.86	64.25	65.14 v	65.19 v	64.63
Tree Size	9-43	9-45	9-39	9-39	9-39

Table 4.67 Experiment 4.52

Findings

The highest CA achieved was 66.62. The tree is performing significantly better than chance and but the absolute CA is poor. The baseline (Conf = 0.25, MNO = 2) range of tree sizes was 81-313. Pruning achieved a very large reduction to 3-11 and the maximally pruned tree had the highest CA.

Discussion

Simulated clauses are intended to investigate how DTs will perform when faced with the most challenging Compound and Complex sentences. *Prima facie*, this experiment provides evidence to reject the null hypothesis and accept that the tree is able to perform some discrimination between complex sentences containing questions and non-questions. However, the CA is relatively low and the DT is simple, suggesting that the DT has not learned much about the domain.

4.12 Comparative analysis of experiments

The purpose of this section is to take an overview of how the results of different groups of experiments relate to each other. A general summary of the best CAs obtained for the experiments in this chapter is shown in table 4.68 below.

Thick lines separate groups of related experiments. Initial findings are that, for alphabetically sorted function words, decision trees make good classifiers for combinations of simple questions with simple non-questions and that they are robust to the presence of preamble inserted in front of both classes. When preambles are inserted in front of only one of the classes, this results in a drop in performance, but the CA is still promising considering that this is the first set of experimental conditions to be tried. When the distracting information is more diverse (mixed length preamble and simulated clauses), performance is affected more severely, however, this may be a side effect of diluting the training set (fewer training examples for each length of preamble or clause). Collectively, these experiments provide evidence that it will be possible to produce classifiers with a high enough CA for real-world applications, after further development.

For all of the experiments in this section, the level of CA that would be obtained by chance is 50% and all of the results reported in the section represent a statistically significant improvement over chance.

Alphabetically Sorted Function Words Experiment	Classification Accuracy	Tree Size
Straightforward QNQ no preambles (4.1-4.4)	98.51	29-67
Straightforward QNQ 1 word preamble (4.9-4.12)	88.11	31-91
Straightforward Q with 2 word preambles / Straightforward NQ (4.13-4.16)	89.62	53-111
Straightforward Q 3-word preambles / Straightforward NQ (4.17-4.20)	88.86	71-131
Straightforward QNQ mixed 1-3 word preambles (4.5-4.8)	79.22	113-189
Straightforward QNQ both with 1 word preambles (4.21-4.24)	98.53	27-29
Straightforward QNQ both with 2 word preambles (4.25-4.28)	98.53	29-73
Straightforward QNQ both with 3 word preambles 4.29-4.32)	98.44	29-33
Straightforward Q difficult NQ no preambles (4.33-4.36)	89.55	49-123
Straightforward Qdifficult NQ both with 1 word preambles (4.37-4.40)	89.93	45-95
Straightforward Q difficult NQ both with 2 word preambles (4.41-4.44)	89.03	53-107
Straightforward Q difficult NQ both with 3 word preambles (4.45-4.48)	89.33	53-121
Simulated clauses (4.49-4.52)	66.62	3-11

Table 4.68 A summary of the experiments involving alphabetically sorted function words

Experiments 4.1-4.4 investigated the simplest scenario for the decision tree to model and consequently set a (very high) upper bound of 98.51 CA to the expectations for following experiments.

Experimental Series	Best C4.5 CA	Best Pruned C4.5 CA
4.1-4.4	98.51	98.36

Table 4.69 Straightforward questions vs. straightforward nonquestions without preambles

Experimental Series	Best C4.5 CA	Best Pruned C4.5 CA
4.5-4.8	79.22	77.87

Table 4.70 Straightforward questions with mixed preambles vs. straightforward nonquestions

This experiment also explored how much the classifier relied on function words (particularly wh-chefts) in the first word position in the sentence to make its classification.

By placing a mix of 1, 2 and 3 word preambles (which themselves could be any combination of function and content words) in front of the straightforward question forms, the task was made more challenging.

Although there was a drop of 19.29 in CA, at 79.22 this DT is still performing much better than chance, providing evidence that the DT has adapted to the more complex feature structure.

Experimental Series	Best C4.5 CA	Best Pruned C4.5 CA
4.9-4.12 (1-word)	88.11	85.46
4.13-4.16 (2-word)	89.62	87.88
4.17-4.20 (3-word)	88.86	87.44

Table 4.71 Straightforward questions with preambles vs. straightforward nonquestions

Experiments 4.9-4.20 investigated the possibility of using a separate DT for each preamble length (on the assumption that these could be used in a multi-classifier system). The results were promising, with the CAs consistently falling about halfway between those for the no preambles and mixed preambles experiments.

Experimental Series	Best C4.5 CA	Best Pruned C4.5 CA
4.21-4.24 (1-word)	98.53	97.66
4.25-4.28 (2-word)	98.53	98.43
4.29-4.32 (3-word)	98.42	98.35

Table 4.72 Straightforward questions vs. straightforward nonquestions both with preambles

Experiments 4.21-4.32 served two purposes. First the series provides a different challenge by introducing more noise at the start of the sentences. Second, it provides a check that the

preambles themselves did not introduce features that aided classification. This was possible because there were a limited number of preambles so each was used several times. Therefore experiments 4.21-4.32 repeat 4.9-4.20 but with the preambles prefixing **both** questions and non-questions. If the preambles were helping the classifier in experiments 4.9-4.20 one would expect a decrease in performance in the corresponding results for 4.21-4.32. In fact the CAs exceed those in 4.9-4.20 and are only marginally below the upper bound to performance established in experiments 4.1-4.4. This provides evidence that the DTs can learn to ignore preambles in finding suitable attributes for decision tree construction.

Experimental Series	Best C4.5 CA	Best Pruned C4.5 CA
4.33-4.36	89.55†	87.77†

Table 4.73 Straightforward questions vs. difficult nonquestions without preambles

Experiments 4.33-4.36 investigate the potential reliance on the first word of the sentence from a different perspective. In this case the difficult non-question has the same kinds of words (particularly wh-chefts) in the first word position as the questions. The CA for this experiment is about 10% lower than the corresponding experiments 4.1-4.4. This provides further evidence that the DTs are not critically dependent on the first word position and can adapt to find more complex features.

Experimental Series	Best C4.5 CA	Best Pruned C4.5 CA
4.37-4.40	89.93†	86.88†
4.41-4.44	89.03†	86.22†
4.45-4.48	89.33†	87.65†

Table 4.74 Straightforward questions vs. difficult nonquestions with preambles

Again experiments 4.37-4.48 investigate whether the DT has compensated for the disruption of the first word position by learning features from the preambles. The outcome of each of the series is highly similar to that for experiments 4.33-4.36. These experiments also correspond to experiments 4.21-4.32 (straightforward questions vs. straightforward nonquestions with preambles). The CAs are about 10% lower than the corresponding

experiments, which is very consistent with the previous experiment. Overall the evidence is that the decision tree has learned to ignore the preambles.

Series	Best C4.5 CA	Best Pruned C4.5 CA
4.49-4.52	66.62†	66.62†

Table 4.75 Straightforward questions vs. straightforward non-questions both preceded by simulated clauses (difficult non-questions)

The final series of experiments in this chapter 4.49-4.52 were intended to simulate a sentence where the question was preceded by a purposeful and substantive clause. To do this both the straightforward questions and the straightforward non-questions were preceded by the set of difficult non-questions (randomly allocated in each case). The results show some promise, with the C4.5 trees achieving statistically significant margins of improvement over the baseline. However, they are too low to be useful in a real-world application.

4.13 General Conclusions

This chapter has presented a new model of semantic similarity and a new method of classifying DAs. The proposed method of classifying DAs has been supported by a series of rigorously designed experiments. These experiments have provided good evidence to support the various hypotheses derived from the proposals that function words can be used to form effective features for question classification and that decision trees will be effective classifiers exploiting those features

Experiments 4.1-4.4 set a high upper bound on the best CA that might be achieved by a decision tree classifier using function word features for question classification. A CA of 98.51% would support very good performance if it could be achieved in real-world systems. Whilst the variants described in the experiments that follow 4.1-4.4 do not achieve this upper-bound, the performance provides a respectable starting point for further investigations into modification of the technique.

Experiments 4.5-4.9 are quite testing because the variable length preambles spread the features around positions where they would be expected in the utterance. Despite this the best CA achieved, 79.22%, is very promising and there is potential for significant improvement with larger training sets.

Various experiments explored a multi-classifier approach in which each preamble length has its own classifier. The most demanding variant in these experiments (4.41-4.44) was the case where the straightforward question form and the difficult non-question form were both preceded by a 2-word preamble (which could contain one or two distracting function words). In this case the best CA achieved was 89.3%.

The experiments on preceding questions and non-questions with clauses produced *prima facie* evidence that the approach taken in this work was successful in classifying them. However, in the pruned C4.5 tree the split is simply on whether the value of W25 is greater or less than a particular value. The significance of W25 may simply be an artefact of generating permutations of two sentences which are 1-25 words long. Despite the statistical significance of the CA of 66.62, it is does not support a claim that experiments 4.49-4.52 provide strong evidence to support the hypothesis.

With the exception of experiments 4.49-4.52, the results reported in this chapter all provide evidence to support the basic alternative hypothesis:

H_1: Function word features can discriminate between questions and non questions.

Furthermore, experiments 4.49-4.52 do not provide evidence to support the null hypothesis.

Therefore the conclusion on the current evidence is to accept the alternative hypothesis that decision trees using function features are effective classifiers of utterances into the question / no question categories.

In some of the later experiments CAs are too low for the DTS to be useful in real-world applications. Therefore, the most important step to take next is to investigate whether the DTs can be optimised. This is explored in chapter 5 by experimenting with different forms of feature encoding.

Chapter 5 Optimising the structure of function word features

5.1 Introduction

The experiments concerning the classification of dialogue acts (chapter 4) made no assumptions about any underlying structure within the class of function words. Consequently they were simply sorted into alphabetical order to create the function word coding table. However, there are known grammatical sub-classes within function words such as pronouns, determiners, link words etc. This chapter explores an approach to optimising function word feature extraction by exploiting the fact that function words can be clustered or grouped using common functional/grammatical properties. Putting function words into such groups or clusters could allow much more effective partitioning of the feature space.

Two variations have been explored. In the **clustering** approach each function word has a unique numeric token, but words are sorted in the coding table so that words belonging to any particular grammatical or functional sub-class are contiguous.

Grouping sorts the function words into the same sub-classes as clustered coding, but in this case each sub-class has a unique code and every member is represented by the same code.

To clarify the above, under the alphabetic sorting approach *I, we* and *you* are numbered 99, 227 and 260 respectively. Under the clustered approach they are 1, 2 and 3 respectively and under the grouped approach they are all allocated the number 1. The significance of grouping data is explained in section 4.6.5.

Knowledge of the English language led to four *a priori* intuitions formulated as hypotheses to be tested in this chapter.

1. Trees trained with either clustered or grouped function words would perform better, in terms of Classification Accuracy (CA) than alphabetically sorted function words.
2. Trees trained with either clustered or grouped function words would perform better, in terms of DT size than alphabetically sorted function words.
3. Trees trained with grouped function words would perform better, in terms of CA than trees trained with clustered function words.
4. Trees trained with grouped function words would perform better, in terms of DT size than trees trained with clustered function words.

These lead to the following sets of general hypotheses:

H_0: The clustering or grouping of function words by functional/grammatical properties has no effect on classification accuracy compared with alphabetically sorted function words.
H_1: The clustering or grouping of function words by functional/grammatical properties increases classification accuracy compared with alphabetically sorted function words.

H_0: The clustering or grouping of function words by functional/grammatical properties has no effect on DT size compared with alphabetically sorted function words.
H_1: The clustering or grouping of function words by functional/grammatical properties reduces DT size compared with alphabetically sorted function words.

H_0: A DT produced by grouping function words will not perform significantly differently from a DT produced by clustering function words, using the same training set.
H_1: A DT produced by grouping function words will have a significantly higher CA than a DT produced by clustering function words, using the same training set.

H_0: The size of a DT produced by grouping function words will not be significantly different from the size of a DT produced by clustering function words, using the same training set.
H_1: The size of a DT produced by grouping function words will be significantly smaller than the size of a DT produced by clustering function words, using the same training set.

5.2 Function word taxonomy

This work has taken the grammatical categories described in comprehensive analyses of the English Language (Carter and McCarthy, 2006, Quirk et al., 1985) and applied them to create a taxonomy of function words through a process of decomposition. The taxonomy forms the basis of the tokenising schemes used for feature extraction and supports the processes of either clustering or grouping the function words.
The taxonomy was derived from the modern view of English grammar, which has in part been informed by the analysis of large corpora (Carter and McCarthy, 2006, Quirk et al., 1985, Sinclair, 2001).
Some simplifications have been made for computational purposes. In particular, where a word could fit into more than one category, it has been allocated to the category associated

with most frequent sense or usage of the word. This process was informed by the Cobuild dictionary (Sinclair, 2001) which ranks the definitions by word sense frequency. The structure of the taxonomy is shown in the following sections. The definitions and examples for the categories are largely derived from the Cobuild dictionary and from Carter & McCarthy.

5.3 Top-level decomposition

Figure 5.1 shows the top level of the decomposition used in this study. A brief description of categories is given below:

- Pronouns

  A pronoun is a word that stands in place of a noun, frequently referring to a noun which has been used earlier in dialogue. Examples include *I, you, she*.
- Determiners

  A determiner is word used before a noun or noun phrase, for example *this, each*.
- Link words

  Link words are in interesting category, because although many can be used as simple adverbs or prepositions, they very often function as introducers of prepositional or adverbial phrases/clauses. Therefore they qualify as function words in this study. Examples include *although, while* and *whenever*.
- Abbreviations

  An abbreviation is a shortened form of a word or phrase, for example *i.e., etc.*
- Prepositions

  A preposition is a word that connects a noun to adjectives or adverbs in making up a phrase, for example *by, at, for*.
- Auxiliary verbs

  An auxiliary is used with a main verb to form different tenses or the passive voice, for example *be, have*. The modal auxiliary verbs are associated with constructions concerning permission or ability, for example *can, may, might*.
- Positives/Negatives

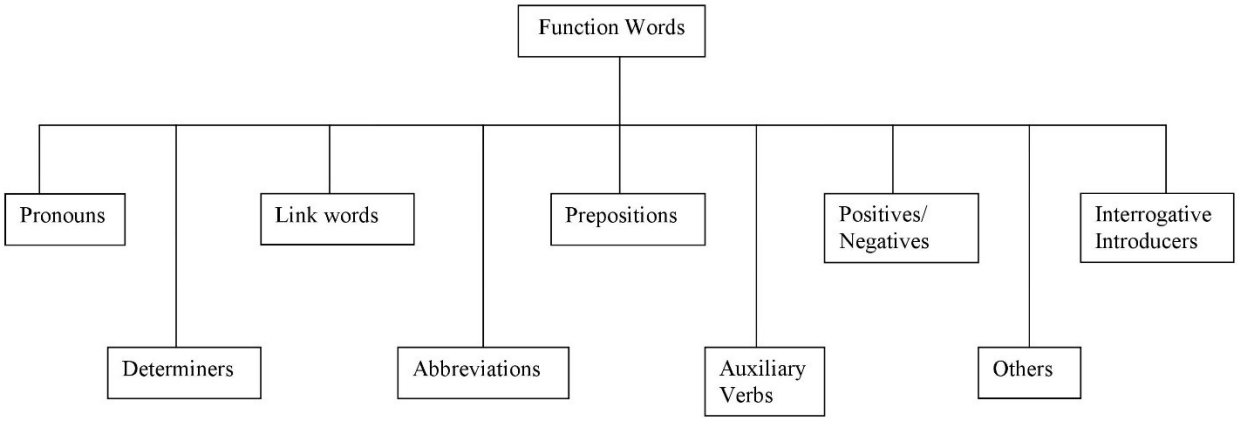
  Positives and negatives are simply words like *yes, no* and *not*.
- Interrogative introducers

  An interrogative introducer is a word used to start a question, for example *where, why, how* (these words are also called wh-chefts)

- Others

  Others are function words which don't fit naturally into any of the previously defined categories, for example *already*, *indeed*.

Figure 5.1 Top level Decomposition of Function Words into Functional Classes



As described in chapter 4, the function words were used to create an encoding table (a 2-D array). Two decisions were made about the order of the groups and words in the table. First, it was decided to put the interrogative introducers (who, how, why etc.) at one end of the table, to maximise the ease of partitioning these words which are so important in forming a question from the others. Second, pairs of classes which might share ownership of words were arranged contiguously and with the ambiguous words at the boundaries. This would allow the DT construction algorithm processing the clustered function words some discretion in deciding which group a word would be placed in when splitting took place.

There is potential for experimenting with alternative orders of the major categories to improve performance. Such improvements, in a decision tree, would arise from being able to partition the feature space in a way that preserved more information for later splits during tree construction.

5.4 Decomposition of the major categories

A number of the top-level categories in figure 5.1 can undergo additional decomposition revealing more detailed structure. Exploiting this is important because if a category contains two different sub-categories each supporting a different final classification, all cases will be classified as belonging to the majority class (with a consequent loss of CA) unless the category is split.

Further decomposition ranges from a relatively simple decomposition of the pronouns to a quite complex structure revealed for the link words.

Some groups: Abbreviations, Prepositions, Positives/Negatives, Interrogative Introducers and Others have not been further decomposed for this study for a variety of reasons including no apparent further structure (Abbreviations, Others), no apparent gain to be made by splitting (Positives/Negatives, Interrogative Introducers) or insufficiently developed theory (Prepositions).

5.4.1 Decomposition of the pronouns

Figure 5.2 shows the decomposition of the pronouns used in this study. A brief description of each of the sub-categories is given below:

- Personal

A personal pronoun is used to refer to the speaker or person spoken to in a dialogue and also to refer to a person or thing which has already been mentioned, for example *I, you, it*.

- Reflexive

  A reflexive pronoun is a word that refers back to the subject of a sentence or clause, for example *myself, himself*.

- Indefinites

  An indefinite pronoun is a word which refers in a general manner to some unspecified person or thing, for example *something, anybody*.

- Posessives

  A possessive pronoun is a word which refers to a thing which belongs to somebody, for example *mine, yours*.

This decomposition was used because it appeared to be the most suitable. Different decompositions are possible, however. For example, clustering the pronouns by person so that *I, mine, myself* etc. would be clustered, whereas *you, yours, yourself* etc. would form part of a different cluster. Such an investigation could form part of future work.

5.4.2 Decomposition of the determiners

Figure 5.3 shows the decomposition of the determiners used in this study. In general terms, a determiner is used to specify which of a number of things is being referred to in a dialogue. A determiner can also specify whether an utterance is referring to one or more things.

Examples for each of the sub-categories are given below:

- Personal

  my, your?

- Numerical

  One, first second, third

- Demonstrative

  This, those

- Definite / Indefinite

  A, an, the

- Others

  All, either

Figure 5.2 Decomposition of the Pronouns Class into Functional Sub-Classes

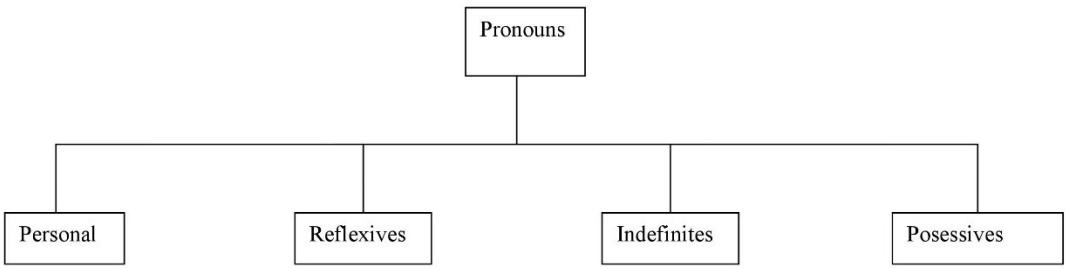
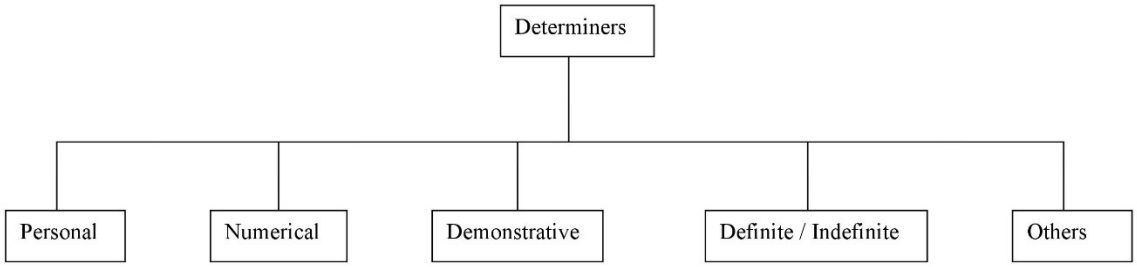


Figure 5.3 Decomposition of the Determiners Class into Functional Sub-Classes



5.4.3 Decomposition of the link words

Figure 5.4 shows the decomposition of the link words used in this study. In general terms a link word is used to connect clauses and sentences.

- Co-ordinators

  And, but

- Subordinators
  - Additives

    Again, also
  - Resultatives

    Hence, so
  - Contrastives

    Else, instead
  - Time indicators

    After, meanwhile
  - Concessives

    Anyhow, anyway
  - Miscellaneous

    Else, as

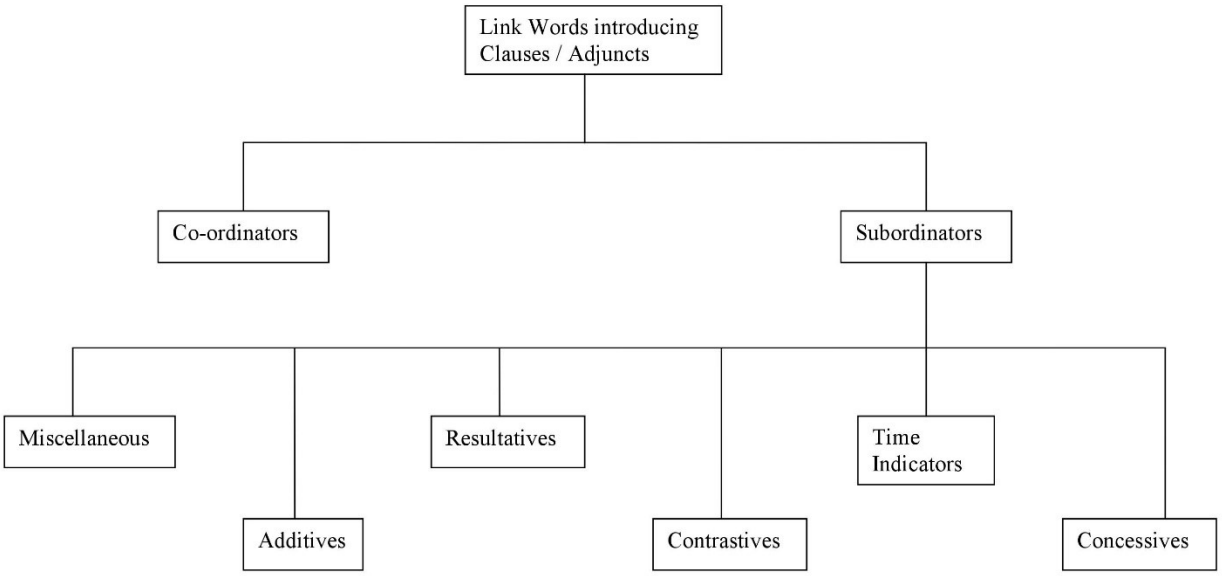
5.4.4 Auxiliaries

In the experiments that follow where each function word is assigned to a unique numeric token, the standard auxiliaries appear first then the modal auxiliaries. As the numbering is contiguous the DT could well discover a useful difference between the two subclasses. In experiments where a group of function words is collectively represented by a single token, the standard and modal auxiliaries form a single group.

5.5 Application of the taxonomy in the new feature encoding method

Once the taxonomy was created, the two new methods of feature encoding were implemented by manually sorting the coding table from chapter 4 into a new order. The order of the function words in the table was the same, but different mappings from words to tokens were used in the two new methods.

Figure 5.4 Decomposition of the Link Words Class into Functional Sub-Classes



5.5.1 Clustered coding

The full coding table for the clustered approach is provided in appendix 5.1. Following this approach each function word in the table has a unique code number. However, instead of being sorted alphabetically the words are grouped by functional / grammatical properties so that words in the group are contiguous. For example, two interrogative introducers, like *how* and *why*, which were widely spaced in the coding table used in chapter 4 (*how* = 97, *why* = 251), are now adjacent (*how* = 275, *why* = 274).

5.5.2 Grouped coding

The full coding table for the grouped approach is provided in appendix 5.2.
Following this approach a single code number (unique to each group) is allocated to all of the words within a functional / grammatical group. For example, all of the auxiliary verbs are allocated the token 193 under this scheme. The rationale for grouping is that a single training case should be able to represent a larger number of feature combinations leading to the same classification. For example, one auxiliary verb / personal pronoun can train the tree to recognise all permutations. So by learning that "Can I" is the start of a question, the tree should also recognise "Will you", "Should they" etc.

5.6 Experimental Methodology

The design of experiments in this chapter follows closely that used in chapter 4. The same procedures are used on a subset of the datasets from chapter 4. This subset consists of the datasets where the classification was more challenging, as there is more room for an improvement in classification accuracy. A description of this subset follows in section 5.6.1.
The difference in the approach taken in this chapter is in the method of encoding the function word features. The first set of experiments uses a version of the function word coding table described as "clustered", the second uses a set described as "grouped". The aim of the experiments is to investigate a series of specific instances of the general hypotheses stated in section 5.1; effectively to investigate whether or not the alternative coding methods produce better DT classifiers in terms of CA and Tree size. Tree size is important for two reasons. First because it is accepted by the DT community that an appropriately pruned DT will generalise better over unseen cases when performing classification (Quinlan, 1986, Quinlan, 1993). Second because the decision trees produced will be more computationally efficient and will scale better on web-based applications with

large numbers of users. Consequently, each experimental series consists of 4 experiments reporting CA and DT size for a range of pruning parameter values.

The first determines the CA and range of DT sizes obtained for pruning using a fixed set of confidence intervals (0.25, 0.2, 0.15, 0.1 and 0.5). This is followed by a second experiment determining CA and DT sizes around the point at which optimum pruning using the confidence interval parameter has been reached. The third experiment determines the CA and range of DT sizes obtained for Minimum Number of Objects (MNO) pruning using a fixed set of minima (2, 5, 10, 15, 20). This is followed by a fourth experiment determining CA and DT sizes around the point at which optimum pruning using the MNO parameter has been reached. In both confidence interval pruning and MNO pruning, the optimal pruning is determined by the smallest tree size that can be achieved before there is a significant reduction in accuracy against a baseline measure. In these experiments, the baseline is a decision tree built using the values of 0.25 for Confidence and 2 for MNO, which are the default settings for the C4.5 version 8 classifier in Weka.

The experiments varying confidence fix MNO at 2 and the experiments varying MNO fix confidence at 0.25 (thus the parameter being varied dominates the pruning).

Occasionally the confidence interval form of pruning can not reach a significant reduction in CA. This is because in very simple domains the value of the parameter becomes so small that floating point arithmetic errors crash the DT construction program before it reaches that point.

In some cases the experimental series did not discover the maximum CA or the optimum pruning level. In these cases more experiments were performed and the specific values of the parameters were reported in addition to the standard experiments.

5.6.1. Selection of experiments

A subset of experiments was selected from those in chapter 4 to represent the range of challenge posed by the different datasets.

1. Straightforward Questions vs. Straightforward Nonquestions without preambles (corresponding to original experiments 4.1-4.4)
2. Straightforward Questions vs. Straightforward Nonquestions with 1-word preambles (corresponding to original experiments 4.9-4.12)
3. Straightforward Questions vs. Straightforward Nonquestions with a mix of 1 – 3 word preambles (corresponding to original experiments 4.5-4.8)

4. Straightforward questions vs. Difficult NonQuestions without preambles (corresponding to original experiments 4.33-4.36)
5. Straightforward questions vs. Difficult NonQuestions both with 2 word preambles (corresponding to original experiments 4.41-4.44)
6. Simulated clauses (corresponding to original experiments 4.49-4.52)
7. Straightforward questions vs. an approximately 50/50 mix of Straightforward and Difficult NonQuestions without preambles used to select the classifier in chapter 4).

Sample 7 represents the dataset used in chapter 4, section 4.5.2 to determine which classifier would be used for the experiments. This dataset was adopted for chapter 5 for 2 reasons. The first was that providing a challenge is more important than investigating particular forms of question and non-question in this chapter. The second was that its lack of balance was not an issue as the datasets used in the grouped form of feature encoding are inherently unbalanced. For consistency of comparison, this dataset was run through the same optimisation experiments as the other datasets in chapter 4 before performing the comparisons in this chapter.

In the following experiments, each run (e.g. experiment 4.1) produces a series of DTs using different parameters. A feature of Weka is that it uses a particular set of parameter values as a baseline and reports the statistical significance of variations from this CA. This is performed using the corrected re-sampled *t*-test.

The important point about this is that all of the experiments were arranged so that baseline trees used for comparison were created using confidence = 0.25 and minimum number of objects = 2, even if these trees do not appear in a table of results.

In the tables of results a significant decrease in CA is marked with a * and a significant increase with the letter v.

Section 5.7 presents a series of experiments using the sample datasets to build DTs using the clustering technique and section 5.8 presents a corresponding series using the grouping technique.

The following experiments are conducted according to the same principles as those for alphabetically sorted function words, described in chapter 4, section 4.10.3.

5.7 Results of experiments using Clustered Coding

The following sets of experiments are presented in groups of 4. Each group of 4 corresponds to one combination of question / non question data listed in table 4.15. The

experiments show how the maximum classification accuracy and the best pruned decision tree size obtainable for each combination was determined, with the function words sorted into clusters according to their common function. Two of the 4 tables from each group show the outcome of confidence interval pruning and two show the results of minimum number of objects pruning.

In the following tables Conf refers to the confidence interval, Min refers to the minimum number of objects allowed in a leaf, CA refers to the percentage classification accuracy and Tree size refers to the number of nodes in a tree.

Experiments 5.1 to 5.4 (straightforward questions / straightforward non-questions, both without preambles)

H_0: Clustering Function Words into functional groups does not improve classification accuracy significantly over alphabetically sorted function words using the data set of straightforward questions vs. straightforward non-questions.

H_1: Clustering Function Words into functional groups improves classification accuracy significantly over alphabetically sorted function words using the data set of straightforward questions vs. straightforward non-questions.

H_0: Clustering Function Words into functional groups does not produce more compact trees than those from alphabetically sorted function words using the data set of straightforward questions vs. straightforward non-questions.

H_1: Clustering Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words using the data set of straightforward questions vs. straightforward non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	98.48	98.55	98.39	98.46	98.65
Tree Size	7-23	7-21	7-21	7-21	7

Table 5.1 Experiment 5.1

Conf	7.0E-8	6.0E-8	5.0E-8	4.0E-8	3.0E-8
CA	98.65	98.65	98.65	98.65	98.65
Tree Size	7	7	7	7	7

Table 5.2 Experiment 5.2

Min	2	5	10	15	20
CA	98.48	98.67	98.65	98.65	98.65
Tree Size	7-23	7-15	7	7	7

Table 5.3 Experiment 5.3

Min	42	43	44	45	46
CA	97.75	97.75	97.65	97.33*	96.85*
Tree Size	7	7	3-9	3-7	3-7

Table 5.4 Experiment 5.4

Findings

The highest CA achieved was 98.67. For the alphabetically sorted Function Words the highest CA was 98.51. The improvement of 0.16 in the CA is not statistically significant, but the reduction in the tree size from 29-67 for the alphabetically sorted function words down to 7-15 (for the best CA in each case) was substantial.

Discussion

Although the improvement in CA is not statistically significant, at this level of CA any improvement is important. The reduction in tree size provides evidence to support the hypothesis that the clustered function words produce more compact DTs by providing better features.

Experiments 5.5 to 5.8 (Straightforward questions with mixed 1-3 word preambles / straightforward non-questions, no preambles)

H_0: Clustering Function Words into functional groups does not improve classification accuracy significantly higher than chance over alphabetically sorted function words using the data set of straightforward questions with 1 – 3 word preambles vs. straightforward non-questions.

H_1: Clustering Function Words into functional groups improves classification accuracy significantly higher than chance over alphabetically sorted function words using the data set of straightforward questions with 1 – 3 word preambles vs. straightforward non-questions.

H_0: Clustering Function Words into functional groups does not produce more compact trees than those from alphabetically sorted function words using the data set of straightforward questions with 1 – 3 word preambles vs. straightforward non-questions.

H_1: Clustering Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words using the data set of straightforward questions with 1 – 3 word preambles vs. straightforward non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	90.70	90.88	90.24	89.86	90.34
Tree Size	87-149	71-129	69-117	65-109	51-89

Table 5.5 Experiment 5.5

Conf	0.01	0.009	0.008	0.007	0.006
CA	89.93	89.88	89.51	89.07*	88.95*
Tree Size	37-71	37-71	37-71	33-71	25-71

Table 5.6 Experiment 5.6

Min	2	5	10	15	20
CA	90.70	89.91	90.46	88.93*	86.11*
Tree Size	87-149	53-81	37-59	31-43	23-39

Table 5.7 Experiment 5.7

Min	10	11	12	13	14
CA	90.46	90.60	89.43	88.77*	88.73*
Tree Size	37-59	37-53	37-53	31-53	31-47

Table 5.8 Experiment 5.8

Findings

The highest CA achieved was 90.88**.** For the alphabetically sorted Function Words the highest CA was 79.22. The improvement of 11.66 is both statistically significant and substantial. Also, the reduction in pruned tree size from 113-189 down to 71-129 (for the best CA in each case) was substantial

Discussion

The findings provide evidence to support the hypothesis that clustering function words provides better features for classification. Furthermore the reduction in tree size provides evidence to support the hypothesis that the clustered function words produce more compact DTs by providing better features.

Experiments 5.9 to 5.12 (Straightforward questions with 1 word preambles / straightforward non-questions, no preambles)

H_0: Clustering Function Words into functional groups does not improve classification accuracy significantly higher than chance over alphabetically sorted function words using the data set of straightforward questions with 1 word preambles vs. straightforward non-questions.
H_1: Clustering Function Words into functional groups improves classification accuracy significantly higher than chance over alphabetically sorted function words using the data set of straightforward questions with 1 word preambles vs. straightforward non-questions.

H_0: Clustering Function Words into functional groups does not produce more compact trees than those from alphabetically sorted function words using the data set of straightforward questions with 1 word preambles vs. straightforward non-questions.
H_1: Clustering Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words using the data set of straightforward questions with 1 word preambles vs. straightforward non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	90.75	91.03	90.44	90.06	90.54
Tree Size	87-135	73-133	65-117	65-109	55-85

Table 5.9 Experiment 5.9

Conf	0.008	0.007	0.006	0.005	0.004
CA	89.88	89.38	89.26	88.33*	87.74*
Tree Size	37-67	33-67	33-65	25-63	25-57

Table 5.10 Experiment 5.10

Min	2	5	10	15	20
CA	90.75	90.03	90.69	88.86*	86.20*
Tree Size	87-135	53-85	37-55	31-43	19-39

Table 5.11 Experiment 5.11

Min	10	11	12	13	14
CA	90.69	90.59	89.43	88.71*	88.63*
Tree Size	37-55	35-53	35-51	33-45	33-45

Table 5.12 Experiment 5.12

Findings

The highest CA achieved was 91.03. For the alphabetically sorted Function Words the highest CA was 88.11. The improvement of 2.92 in the CA was statistically significant, however, the increase in pruned tree size from 31-91 to 73-133 (for the best CA in each case) was substantial.

Discussion

The increase in CA provides evidence to support the hypothesis that clustering function words provides better features for classification. Unfortunately the increase in tree size appears to contradicts the hypothesis that the clustered function words produce more compact trees. Looking a little more closely, clustered function words produced more compact trees than the alphabetically sorted function words for equivalent CA (88.33% CA ranges from 25-63 nodes) using this dataset. Therefore there is no evidence that clustering produces inherently larger trees.

Experiments 5.13 to 5.16 (Straightforward questions / difficult non-questions, both without preambles)

H_0: Clustering Function Words into functional groups does not improve classification accuracy over alphabetically sorted function words with straightforward questions vs. difficult non-questions.

H_1: Clustering Function Words into functional groups improves classification accuracy over alphabetically sorted function words with straightforward questions vs. difficult non-questions..

H_0: Clustering Function Words into functional groups does not produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. difficult non-questions..

H_1: Clustering Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. difficult non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	91.58	91.24	91.10	91.49	91.72
Tree Size	61-105	57-99	49-99	35-77	25-69

Table 5.13 Experiment 5.13

Conf	7.0E-8	6.0E-8	5.0E-8	4.0E-8	3.0E-8
CA	90.30	90.30	90.30	90.30	90.30
Tree Size	9-17	9-17	9-17	9-15	9-15

Table 5.14 Experiment 5.14

Min	2	5	10	15	20
CA	91.58	90.88	90.16	91.35	90.80
Tree Size	61-105	39-63	19-39	13-27	13-23

Table 5.15 Experiment 5.15

Min	25	26	27	28	29
CA	90.69	90.37	90.00	89.67*	89.63*
Tree Size	9-23	9-23	9-15	9-15	9-15

Table 5.16 Experiment 5.16

Findings

The highest CA achieved was 91.72**.** For alphabetically sorted function words, the highest CA was 89.55. The improvement of 2.17 in the CA is statistically significant. Additionally, the reduction in pruned tree size from 49-123 down to 25-69 (for the best CA in each case) was substantial.

Discussion

The increase in CA provides evidence to support the hypothesis that clustering function words provides better features for classification. Also, the reduction in tree size provides evidence to support the hypothesis that the clustered function words produce more compact trees by providing better features.

Experiments 5.17 to 5.20 (Straightforward questions / difficult non-questions, both with 2 word preambles)

H_0: Clustering Function Words into functional groups does not improve classification accuracy over alphabetically sorted function words with straightforward questions vs. difficult non-questions when both have 2 word preambles.

H_1: Clustering Function Words into functional groups improves classification accuracy over alphabetically sorted function words with straightforward questions vs. difficult non-questions when both have 2 word preambles.

H_0: Clustering Function Words into functional groups does not produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. difficult non-questions when both have 2 word preambles.
H_1: Clustering Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. difficult non-questions when both have 2 word preambles.

Conf	0.25	0.2	0.15	0.1	0.05
CA	91.13	90.91	90.82	91.32	91.53
Tree Size	75-115	63-107	55-105	35-79	29-67

Table 5.17 Experiment 5.17

Conf	7E-8	6E-8	5E-8	4E-8	3E-8
CA	90.39	90.39	90.39	90.39	90.39
Tree Size	9-17	9-17	9-17	9-15	9-15

Table 5.18 Experiment 5.18

Min	2	5	10	15	20
CA	91.13	90.83	90.28	91.63	90.98
Tree Size	75-115	41-65	23-37	19-27	13-23

Table 5.19 Experiment 5.19

Min	60	61	62	63	64
CA	89.97	89.88	89.58	88.92*	87.81*
Tree Size	7	7	7	7	7

Table 5.20 Experiment 5.20

Findings

The highest CA achieved was 91.63**.** For alphabetically sorted function words, the highest CA was 89.03. The improvement of 2.60 in the CA is statistically significant and the reduction in pruned tree size from 53-107 down to 19-27 (for the best CA in each case) was substantial.

Discussion

The increase in CA provides evidence to support the hypothesis that clustering function words provides better features for classification. Moreover the reduction in tree size provides evidence to support the hypothesis that the clustered function words produce more compact trees by providing better features.

Experiments 5.21 to 5.24 Simulated Clauses (difficult non-questions inserted in front of straightforward questions / straightforward non-questions)

H_0: Clustering Function Words into functional groups does not improve classification accuracy over alphabetically sorted function words with straightforward questions vs. straightforward non-questions when both are preceded by difficult simulated clauses.
H_1: Clustering Function Words into functional groups improves classification accuracy over alphabetically sorted function words with straightforward questions vs. straightforward non-questions when both are preceded by difficult simulated clauses.

H_0: Clustering Function Words into functional groups does not produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. straightforward non-questions when both are preceded by difficult simulated clauses.
H_1: Clustering Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. straightforward non-questions when both are preceded by difficult simulated clauses.

Conf	0.25	0.2	0.15	0.1	0.05
CA	73.54	73.91	73.94	74.62	75.04v
Tree Size	103-267	83-229	71-211	59-175	51-155

Table 5.21 Experiment 5.21

Conf	1E-5	9E-6	8E-6	7E-6	6E-6
CA	71.38	71.04	70.59	69.98*	69.58*
Tree Size	3-27	3-27	3-27	3-27	3-27

Table 5.22 Experiment 5.22

Min	2	5	10	15	20
CA	73.54	74.46	75.62	75.72	72.76
Tree Size	103-267	69-151	35-87	27-63	23-53

Table 5.23 Experiment 5.23

Min	46	47	48	49	50
CA	72.05	71.13	70.63	70.46*	69.82*
Tree Size	15-27	15-31	15-29	15-29	15-29

Table 5.24 Experiment 5.24

Findings

The highest CA achieved was 75.72%. For alphabetically sorted function words, the highest CA was 66.62. The improvement of 9.1 in the CA is statistically significant.
On the other hand, the increase in pruned tree size from 3-11 to 27-63 (for the best CA in each case) was substantial.

Discussion

The increase in CA provides evidence to support the hypothesis that clustering function words provides better features for classification.
Prima facie, the increase in pruned tree size apparently contradicts the hypothesis that the clustered function words produce more compact trees. But this technique produces similar trees to the alphabetically sorted function words for equivalent CA (69.58%) ranging from 3-27 nodes. The inference is that the dataset is demanding and the alphabetically sorted features only reveal enough detail about the feature space for a simplistic and relatively low performance DT to be built. In this case, clustering has revealed more detail allowing a larger and better performing tree to be constructed.

5.7.1. Summary of Clustered Function Words Experiments

Corresponding Alphabetically Sorted Function Words Experiment	Classification Accuracy	Tree Size
Straightforward QNQ no preambles (4.1-4.4)	98.67	7-15
Straightforward QNQ 1 word preamble (4.9-4.12)	91.03	73-133†
Straightforward QNQ mixed 1-3 word preambles (4.5-4.8)	90.88	71-129
Straightforward Q difficult NQ no preambles (4.33-4.36)	91.72	25-69
Simple Q Difficult NQ both with 2 word preambles (4.41-4.44)	91.63	19-27
Simulated clauses (4.49-4.52)	75.72	27-63

Table 5.25 Summary of the clustered coding experiments

This selection of some of the more challenging datasets from the experiments in chapter 4 shows a clear improvement in CA over the alphabetically sorted function words as a result of the clustered coding (table 5.25). In several cases increases of over 10% have been

achieved and the classifiers are approaching the point at which they could be used in real-world CAs. These results will be discussed in more detail when they are compared with the results of grouped coding.

5.8 Results of experiments using Grouped Coding

The following sets of experiments are presented in groups of 4. Each group of 4 corresponds to one combination of question / non question data listed in table 4.15. The experiments show how the maximum classification accuracy and the best pruned decision tree size obtainable for each combination was determined, with the function words grouped according to function and all of member of a group assigned the same numeric token. Two of the 4 tables from each group show the outcome of confidence interval pruning and two show the results of minimum number of objects pruning.
In the following tables Conf refers to the confidence interval, Min refers to the minimum number of objects allowed in a leaf, CA refers to the percentage classification accuracy and Tree size refers to the number of nodes in a tree.

Experiments 5.25 to 5.28 (straightforward questions / straightforward non-questions, both without preambles)

H_0: Grouping Function Words into functional groups does not improve classification accuracy significantly higher than chance, over alphabetically sorted function words using the data set of straightforward questions vs. straightforward non-questions.
H_1: Grouping Function Words into functional groups improves classification accuracy significantly higher than chance, over alphabetically sorted function words using the data set of straightforward questions vs. straightforward non-questions.

H_0: Grouping Function Words into functional groups does not produce more compact trees than those from alphabetically sorted function words using the data set of straightforward questions vs. straightforward non-questions.
H_1: Grouping Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words using the data set of straightforward questions vs. straightforward non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	98.77	98.77	98.77	98.77	98.77
Tree Size	7	7	7	7	7

Table 5.26 Experiment 5.25

Conf	7.0E-8	6.0E-8	5.0E-8	4.0E-8	3.0E-8
CA	98.77	98.77	98.77	98.77	98.77
Tree Size	7	7	7	7	7

Table 5.27 Experiment 5.26

Min	2	5	10	15	20
CA	98.77	98.77	98.77	98.77	98.77
Tree Size	7	7	7	7	7

Table 5.28 Experiment 5.27

Min	28	29	30	31	32
CA	98.77	98.77	98.45	97.76*	95.95*
Tree Size	7	7	3-7	3-7	3-7

Table 5.29 Experiment 5.28

Findings

In this experiment, after tokenising, the relative proportions of the records in the dataset were 538 Questions, 597 Non-questions and the classification accuracy achieved simply by selecting the majority class would be 52.6.

The highest CA achieved was 98.77**.** For the alphabetically sorted Function Words the highest CA was 98.51 and for the clustered function words it was 98.67. The improvements of 0.26 in the CA over the alphabetically sorted function words and 0.1 over the clustered function words were not statistically significant. Nevertheless, the reduction in pruned tree size from 29-67 for the alphabetically sorted function words and 7-15 for clustered function words down to 7 (for the best CA in each case) was substantial.

Discussion

Even though the improvement in CA is very small, once this level of CA has been reached any improvement is important. In addition the reduction in tree size provides evidence to support the hypothesis that the grouped function words produce more compact DTs by providing better features.

Experiments 5.29 to 5.32 (Straightforward questions with mixed 1-3 word preambles / straightforward non-questions, no preambles)

H_0: Grouping Function Words into functional groups does not improve classification accuracy significantly higher than chance over alphabetically sorted function words using the data set of straightforward questions with 1 – 3 word preambles vs. straightforward non-questions.
H_1: Grouping Function Words into functional groups improves classification accuracy significantly higher than chance over alphabetically sorted function words using the data set of straightforward questions with 1 – 3 word preambles vs. straightforward non-questions.

H_0: Grouping Function Words into functional groups does not produce more compact trees than those from alphabetically sorted function words using the data set of straightforward questions with 1 – 3 word preambles vs. straightforward non-questions.
H_1: Grouping Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words using the data set of straightforward questions with 1 – 3 word preambles vs. straightforward non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	87.07	87.17	87.12	87.21	87.42
Tree Size	97-151	79-133	69-125	57-117	45-99

Table 5.30 Experiment 5.29

Conf	0.001	0.0009	0.0008	0.0007	0.0006
CA	85.14	85.13	85.07*	84.65*	84.56*
Tree Size	19-59	19-59	19-59	17-59	17-59

Table 5.31 Experiment 5.30

Min	2	5	10	15	20
CA	87.07	87.64	87.53	86.96	86.91
Tree Size	97-151	53-95	35-61	27-43	25-33

Table 5.32 Experiment 5.31

Min	30	31	32	33	34
CA	85.37	85.36	85.48	84.48*	83.84*
Tree Size	19-29	17-29	17-29	15-29	15-27

Table 5.33 Experiment 5.32

Findings

In this experiment, after tokenising, the relative proportions of the records in the dataset were Questions 593, Non-questions 597 and the classification accuracy achieved simply by selecting the majority class would be 50.2%.

The highest CA achieved was 87.64**.** For the alphabetically sorted Function Words the highest CA was 79.22 and for clustered function words it was 90.88. The improvement of 8.42 for grouping over alphabetically sorted function words is statistically significant. There was a reduction in CA of 3.24 compared with clustered function words and this was statistically significant, providing evidence that grouping does not improve CA over clustering. Also, the reduction in pruned tree size from 113-189 for the alphabetically sorted function words and 71-129 for clustered function words down to 53-95 (for the best CA in each case) was substantial

Discussion

The increase in CA provides evidence to support the hypothesis that grouping function words provides better features than alphabetically sorted function words for classification. The experiment provides evidence to contradict the hypothesis that grouped function words improve upon alphabetically sorted function words, and this is statistically significant.

There is evidence to support the hypothesis that the grouped function words produce more compact DTs. However, it is not possible to assume the features are better because of the drop in CA.

Experiments 5.33 to 5.36 (Straightforward questions with 1 word preambles / straightforward non-questions, no preambles)

H_0: Grouping Function Words into functional groups does not improve classification accuracy significantly higher than chance over alphabetically sorted function words using the data set of straightforward questions with 1 word preambles vs. straightforward non-questions.

H_1: Grouping Function Words into functional groups improves classification accuracy significantly higher than chance over alphabetically sorted function words using the data set of straightforward questions with 1 word preambles vs. straightforward non-questions.

H_0: Grouping Function Words into functional groups does not produce more compact trees than those from alphabetically sorted function words using the data set of straightforward questions with 1 word preambles vs. straightforward non-questions.
H_1: Grouping Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words using the data set of straightforward questions with 1 word preambles vs. straightforward non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	87.06	87.39	87.52	87.63	87.96
Tree Size	97-151	85-137	75-115	61-109	41-97

Table 5.34 Experiment 5.33

Conf	0.001	0.0009	0.0008	0.0007	0.0006
CA	85.38	85.35	85.29	84.91*	84.84*
Tree Size	17-61	17-61	17-61	13-61	13-59

Table 5.35 Experiment 5.34

Min	2	5	10	15	20
CA	87.06	87.61	87.15	86.98	87.11
Tree Size	97-151	57-95	35-61	27-43	25-33

Table 5.36 Experiment 5.35

Min	30	31	32	33	34
CA	85.45	85.50	85.31	84.54*	84.32*
Tree Size	19-29	17-29	17-29	15-29	15-27

Table 5.37 Experiment 5.36

Findings

In this experiment, after tokenising, the relative proportions of the records in the dataset were 593 Questions, 597 Non-questions and the classification accuracy achieved simply by selecting the majority class would be 50.2%.
The highest CA achieved, 88.47 was found by searching outside the normal range of confidence levels (conf = 0.009). For the alphabetically sorted function words the highest CA was 88.11 and for clustered function words it was 91.03. The improvement of 0.36 in

the CA over the alphabetically sorted function words is not statistically significant, furthermore, there was a reduction in CA of 2.56 compared with clustered function words and this was statistically significant.

The range of tree sizes for the CA of 88.47 was 31-91. In comparison the alphabetically sorted function words had a range of 33-77 for the best CA and clustered function words had a range of 73-133.

Discussion

This experiment does not provide evidence to support the hypothesis that grouping function words provides better features for classification than alphabetically sorted function words. Also, there is significant evidence to contradict the hypothesis that grouped function words produce better classifiers than clustered function words.

So the evidence to support the hypothesis that grouping produces smaller DTs than alphabetic sorting is inconclusive.

Because the range of tree sizes for grouped function words overlaps the range for alphabetically sorted function words it is not possible to be conclusive about which approach is better. Clearly the clustered approach has produced larger tree sizes as the price for its improved CA.

Experiments 5.37 to 5.40 (Straightforward questions / difficult non-questions, both without preambles)

H_0: Grouping Function Words into functional groups does not improve classification accuracy over alphabetically sorted function words with straightforward questions vs. difficult non-questions.

H_1: Grouping Function Words into functional groups improves classification accuracy over alphabetically sorted function words with straightforward questions vs. difficult non-questions.

H_0: Grouping Function Words into functional groups does not produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. difficult non-questions.

H_1: Grouping Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. difficult non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	92.47	92.73	92.27	92.66	92.55
Tree Size	41-91	41-69	25-67	21-65	21-37

Table 5.38 Experiment 5.37

Conf	8.0E-4	70E-4	60E-4	50E-4	40E-4
CA	91.30	91.30	91.24	91.05*	91.05*
Tree Size	9-25	9-25	9-23	9-23	9-23

Table 5.39 Experiment 5.38

Min	2	5	10	15	20
CA	92.47	90.61*	90.95*	92.05	91.34
Tree Size	41-91	21-53	15-31	13-23	13-15

Table 5.40 Experiment 5.39

Min	21	22	23	24	25
CA	91.35	91.33	91.11	90.67*	90.25*
Tree Size	13-15	13-15	11-15	11-15	9-15

Table 5.41 Experiment 5.40

Findings

In this experiment, after tokenising, the relative proportions of the records in the dataset were 538 Questions, 597 Non-questions and the classification accuracy achieved simply by selecting the majority class would be 50.2%.

The highest CA achieved was 92.73. For alphabetically sorted function words, the highest CA was 89.55 and for clustered function words the highest CA was 91.72.

The improvement of 3.18 in the CA is statistically significant. There was a reduction in CA of 1.1 compared with clustered function words which is not statistically significant but at this level of CA is still important.

The reduction in pruned tree size from 49-123 for the alphabetically sorted function words to 41-69 (for the best CA in each case) is substantial. Clustered function words produced even more compact tree with sizes in the range 25-69

Discussion

The experiment provides evidence to support the hypothesis that grouping function words provides better features than alphabetically sorted for classification. The reduction in CA compared with clustered function words provides weak evidence to contradict the hypothesis that grouping function provides better features for classification than clustering.

Again, when comparing the ranges of tree sizes for the different approaches, the results suggest that the grouped function words produce more compact trees and are therefore finding better features.

Experiments 5.41 to 5.44 (Straightforward questions / difficult non-questions, both with 2 word preambles)

H_0: Grouping Function Words into functional groups does not improve classification accuracy over alphabetically sorted function words with straightforward questions vs. difficult non-questions when both have 2 word preambles.
H_1: Grouping Function Words into functional groups improves classification accuracy over alphabetically sorted function words with straightforward questions vs. difficult non-questions when both have 2 word preambles.

H_0: Grouping Function Words into functional groups does not produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. difficult non-questions when both have 2 word preambles.
H_1: Grouping Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. difficult non-questions when both have 2 word preambles.

Conf	0.25	0.2	0.15	0.1	0.05
CA	92.37	92.62	92.05	92.33	92.35
Tree Size	51-101	45-89	29-79	23-67	19-43

Table 5.42 Experiment 5.41

Conf	0.0004	0.0003	0.0002	0.0001	0.00009
CA	91.71	91.10	91.01	90.91*	90.90*
Tree Size	13-25	11-25	11-23	11-23	11-23

Table 5.43 Experiment 5.42

Min	2	5	10	15	20
CA	92.37	90.81*	91.26	93.04	92.30
Tree Size	51-101	23-61	15-27	15-23	15-23

Table 5.44 Experiment 5.43

Min	26	27	28	29	30
CA	92.32	92.29	92.01	91.16	89.79*
Tree Size	13-15	13-15	9-15	9-15	9-15

Table 5.45 Experiment 5.44

Findings

In this experiment, after tokenising, the relative proportions of the records in the dataset were 598 Questions, 599 Non-questions and the classification accuracy achieved simply by selecting the majority class would be 50.1%.

The highest CA achieved was 93.04**.** For alphabetically sorted function words, the highest CA was 89.03 and for clustered function words the highest CA was 91.63.

The improvement of 4.01 in the CA is statistically significant, however the improvement in CA of grouped function words over clustered function words of 1.41, is not statistically significant.

Discussion

The experiment provides evidence to support the hypothesis that grouping function words provides better features than alphabetically sorted function words for classification. The reduction in pruned tree size from 53-107 for alphabetically sorted function words down to 15-23 (for the best CA in each case) was substantial Similarly there is weak evidence for an improvement over clustered function words because the improvement is not statistically significant, but at this level of CA it is still important. Moreover the reduction from the size range for the best CA for clustered function words (19-27), is also an important improvement.

Collectively these results provide strong evidence to support the hypothesis that grouping function words produces more compact trees.

Experiments 5.45 to 5.48 Simulated Clauses (difficult non-questions inserted in front of straightforward questions / straightforward non-questions)

H_0: Grouping Function Words into functional groups does not improve classification accuracy over alphabetically sorted function words with straightforward questions vs. straightforward non-questions when both are preceded by difficult simulated clauses.

H_1: Grouping Function Words into functional groups improves classification accuracy over alphabetically sorted function words with straightforward questions vs. straightforward non-questions when both are preceded by difficult simulated clauses.

H_0: Grouping Function Words into functional groups does not produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. straightforward non-questions when both are preceded by difficult simulated clauses.

H_1: Grouping Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. straightforward non-questions when both are preceded by difficult simulated clauses.

Conf	0.25	0.2	0.15	0.1	0.05
CA	74.34	75.49v	75.94v	76.12v	77.38v
Tree Size	97-257	85-227	77-191	67-165	43-117

Table 5.46 Experiment 5.45

Conf	7E-8	6E-8	5E-8	4E-8	3E-8
CA	75.85	75.64	75.71	75.76	75.75
Tree Size	17-31	17-31	17-31	17-27	17-27

Table 5.47 Experiment 5.46

Min	2	5	10	15	20
CA	74.34	76.53v	76.34	75.17	76.70v
Tree Size	97-257	57-143	33-85	27-55	25-49

Table 5.48 Experiment 5.47

Min	51	52	53	54	55
CA	72.13	72.07	71.78	71.31*	71.09*
Tree Size	13-21	13-21	13-21	13-21	13-21

Table 5.49 Experiment 5.48

Findings

In this experiment, after tokenising, the relative proportions of the records in the dataset were 600 Questions, 600 Non-questions and the classification accuracy achieved simply by selecting the majority class would be 50.

The highest CA achieved 78.84 was found by searching outside the normal range of Minimum Number of Objects levels (MNO = 27). For the alphabetically sorted function words, the highest CA was 66.62 and for clustered function words the highest CA was 75.72.

The improvement of 12.22 over the alphabetically sorted function words was statistically significant and the improvement in CA of the grouped function words of 3.12 over the clustered function words was also statistically significant.

The increase in pruned tree size from 3-11 for the alphabetically sorted function words to 25-35 (for the best CA in each case) was substantial, nevertheless it does show an improvement compared with the range of 27-63 for clustered function words.

Discussion

The results of the experiment provides evidence to support the hypotheses that grouping function words provides better features for classification over both alphabetically sorted and clustered function words.

The increase in the range of tree sizes over alphabetically sorted function words apparently contradicts the hypothesis that the grouped function words produce more compact trees. But they still improve on clustered function words and, as discussed earlier in this section (experiments 5.21 to 5.24) the explanation may simply be that clustering has revealed more detail allowing a larger and better performing tree to be constructed.

5.8.1 Summary of Grouped Function Words Experiments

Corresponding Alphabetically Sorted Function Words Experiment	Classification Accuracy	Tree Size
Straightforward QNQ no preambles (4.1-4.4)	98.77	7
Straightforward QNQ 1 word preamble (4.9-4.12)	88.47	33-77
Straightforward QNQ mixed 1-3 word preambles (4.5-4.8)	87.64	53-95
Straightforward Q difficult NQ no preambles (4.33-4.36)	92.73	41-69
Simple Q Difficult NQ both with 2 word preambles (4.41-4.44)	93.04	15-23
Simulated clauses (4.49-4.52)	78.84	25-35

Table 5.50 Summary of the grouped coding experiments

Again, the selection of some of the more challenging datasets from the experiments in chapter 4 shows a clear improvement in CA over the alphabetically sorted function words as a result of the clustered coding (see table 5.50). The overall improvement is approximately the same as for clustered function words, but with minor variations between the groups of experiments. These results will be discussed in more detail when they are compared with the results of clustered coding, later in this chapter.

5.9 Supplementary Experiments

The following set of experiments uses the preliminary dataset which was only used in chapter 4 (experiment 4.0) to select a classifier. All of the experiments in this chapter are compared with a corresponding set performed in chapter 4 (using alphabetically sorted function words). Because the necessary experiments were not performed in chapter 4, they are performed here as experiments 5.49 -5.52.

The following sets of experiments are presented in groups of 4. Each group of 4 corresponds to the same dataset, straightforward questions / mixed straightforward and difficult non-questions, but in each case with a different tokenising table for the function words; alphabetic, clustered or grouped.

The experiments show how the maximum classification accuracy and the best pruned decision tree size obtainable for each combination was determined, for each of the ways of organising the function words. Two of the 4 tables from each group show the outcome of confidence interval pruning and two show the results of minimum number of objects pruning.

In the following tables Conf refers to the confidence interval, Min refers to the minimum number of objects allowed in a leaf, CA refers to the percentage classification accuracy and Tree size refers to the number of nodes in a tree.

Experiments 5.49 to 5.52 (Straightforward questions / mixed straightforward and difficult non-questions, all without preambles) alphabetically sorted function words

These experiments test the following hypotheses which are specific instances of the general hypotheses stated in chapter 4, providing a basis for comparison needed by the experiments that follow them (5.53-5.56 and 5.57-5.60).

H_0: A DT using function words can not achieve classification accuracy significantly higher than chance over the data set of straightforward questions vs. a mixture of straightforward and difficult non-questions.

H_1: A DT using function words can achieve classification accuracy significantly higher than chance over the data set of straightforward questions vs. a mixture of straightforward and difficult non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	88.73	89.31v	89.32	89.35	89.37
Tree Size	89-147	79-135	79-131	69-117	59-101

Table 5.51 Experiment 5.49

Conf	0.0005	0.0004	0.0003	0.0002	0.0001
CA	88.48	88.25	87.74	85.29*	84.21*
Tree Size	31-65	31-65	27-65	19-61	27-37

Table 5.52 Experiment 5.50

Min	2	5	10	15	20
CA	88.73	88.97	88.29	85.83*	83.64*
Tree Size	89-147	59-99	45-65	31-57	27-45

Table 5.53 Experiment 5.51

Min	11	12	13	14	15
CA	87.89	87.49	86.83	85.70*	85.83*
Tree Size	39-63	39-59	35-59	33-61	31-57

Table 5.54 Experiment 5.52

Findings

In this experiment, after tokenising, the relative proportions of the records in the dataset were 532 Questions, 612 Non-questions and the classification accuracy achieved simply by selecting the majority class would be 53.5 %. The highest CA achieved was 89.37, which is statistically significant and exceeds the majority class level by a large margin.

The baseline range of tree sizes is 89-147. Optimal DTs had sizes in the range 59-101 and reduced range of 35-59 before a significant reduction in CA.

Discussion

The experiments support the hypothesis that function words provide effective features for classification of questions vs. a more demanding combination of non-questions. This combination is challenging because it can not really on obvious features like the presence of a wh-cheft to indicate a question. Nor can it simple ignore the first word as there is still a higher probability that features such as a wh-cheft in the first word will indicate a question. Also the mix of simple and difficult non-questions dilutes the non-question data as a source of features. Consequently, the large and statistically significant improvement over chance levels of classification is good evidence to reject the null hypothesis and

accept that the tree is classifying effectively, modelling the data well, even bearing in mind that it is not as high as would be desirable.

Experiments 5.53 to 5.56 (Straightforward questions / mixed straightforward and difficult non-questions, all without preambles) – clustered function words

H_0: Clustering Function Words into functional groups does not improve classification accuracy over alphabetically sorted function words with straightforward questions vs. a mixture of straightforward and difficult non-questions.
H_1: Clustering Function Words into functional groups improves classification accuracy over alphabetically sorted function words with straightforward questions vs. a mixture of straightforward and difficult non-questions.

H_0: Clustering Function Words into functional groups does not produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. a mixture of straightforward and difficult non-questions.
H_1: Clustering Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. a mixture of straightforward and difficult non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	92.64	92.78	92.74	92.55	92.43
Tree Size	47-85	45-77	37-69	33-59	29-55

Table 5.55 Experiment 5.53

Conf	0.0004	0.0003	0.0002	0.0001	9E-5
CA	92.43	91.96	91.81	91.07*	91.06*
Tree Size	15-27	13-27	13-27	11-25	11-25

Table 5.56 Experiment 5.54

Min	2	5	10	15	20
CA	92.64	93.06	90.08*	90.32*	92.06
Tree Size	47-85	31-59	13-33	13-23	13-21

Table 5.57 Experiment 5.55

Min	6	7	8	9	10
CA	93.14	92.42	92.09	91.03*	90.08*
Tree Size	27-47	27-41	19-35	19-33	13-33

Table 5.58 Experiment 5.56

Findings

The highest CA achieved was 93.14**.** For the alphabetically sorted Function Words the highest CA was 89.37. The improvement of 3.77 in the CA is statistically significant and the reduction in pruned tree size from 59-101 for alphabetically sorted function words down to 27-47 (for the best CA in each case) is substantial.

Discussion

These results provide evidence to support the hypothesis that clustering function words provides better features for classification and also better modelling because the trees are more compact. The increased CA provides some evidence that the better modelling capability is compensating for the sparseness of the training data.

Experiments 5.57 to 5.60 (Straightforward questions / mixed straightforward and difficult non-questions, all without preambles) grouped function words

H_0: Grouping Function Words into functional groups does not improve classification accuracy over alphabetically sorted function words with straightforward questions vs. a mixture of straightforward and difficult non-questions.

H_1: Grouping Function Words into functional groups improves classification accuracy over alphabetically sorted function words with straightforward questions vs. a mixture of straightforward and difficult non-questions.

H_0: Grouping Function Words into functional groups does not produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. a mixture of straightforward and difficult non-questions.

H_1: Grouping Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words with straightforward questions vs. a mixture of straightforward and difficult non-questions.

Conf	0.25	0.2	0.15	0.1	0.05
CA	93.27	93.63	93.43	93.22	92.19*
Tree Size	55-77	51-73	45-69	39-67	33-59

Table 5.59 Experiment 5.57

Conf	0.08	0.07	0.06	0.05	0.04
CA	93.19	93.08	93.03	92.19*	91.90*
Tree Size	37-67	29-63	31-63	33-59	23-55

Table 5.60 Experiment 5.58

Min	2	5	10	15	20
CA	93.27	92.67	91.15*	91.18*	91.78*
Tree Size	55-77	35-53	17-33	13-23	13-15

Table 5.61 Experiment 5.59

Min	6	7	8	9	10
CA	93.35	92.39	91.93	91.91	91.15*
Tree Size	37-49	31-51	23-43	21-39	17-33

Table 5.62 Experiment 5.60

Findings

The highest CA achieved was 93.63. For alphabetically sorted function words, the highest CA was 89.37. The improvement of 4.26 in the CA is statistically significant providing evidence to support the hypothesis that grouping function words provides better features for classification. Clustered function words produced a CA of 93.14.

The improvement in CA of the grouped function words of 0.49 over the clustered function words was not statistically significant.

The reduction in tree size from 59-101 for alphabetically sorted function words down to 51-73 (for the best CA in each case) is substantial, however, the range of tree sizes is larger than that of the clustered function words (27-47).

Discussion

Although the improvement in CA is not statistically significant, as the CA is over 90% it may still be seen as important.

The reduction in tree size compare with alphabetically sorted function words provides evidence to support the hypothesis that the grouped function words produce more compact DTs by providing better features. The fact that the trees are larger than those for the

clustered approach suggests that clustering reveals additional structure in the features, but exploiting this structure requires a larger tree.

The following sections analyse the results of the experiments in this chapter. To aid this, an overall view of the results and corresponding results using alphabetically sorted function words (from chapter 4) is provided in section 5.10, table 5.64.

5.9.1 Summary of mixed Simple /Difficult Non-questions

Function word organisation	Classification Accuracy	Tree Size
Alphabetically sorted	89.37	59-101
Functionally Clustered	93.14	27-47
Functionally Grouped	93.63	49-73

Table 5.63 A comparison of all 3 methods of tokenising function words using the data set Straightforward Questions with mixed Simple /Difficult Non-questions (all without preambles)

The initial observation that can be made is that for this more demanding mix of non-questions, the two methods of organising the function words presented in this chapter produce a substantially higher CA than the alphabetically sorted function words (table 5.63). The performance is close to that which would be suitable for deployment in a real-world CA. A more detailed comparison, including the significance of results follows in sections 5.10 to 5.12.

5.10 Overview of experimental results

A summary of the results of the experiments conducted in this chapter and those from chapter 4 is shown in table 5.64, below:

	Alpha Sorted Function Words		Clustered Function Words		Grouped Function Words	
Clustered Function Words Experiment	Classification Accuracy	Tree Size	Classification Accuracy	Tree Size	Classification Accuracy	Tree Size
Straightforward QNQ no preambles (4.1-4.4)	98.51	29-67	98.67	7-15	98.77	7
Straightforward QNQ 1 word preamble (4.9-4.12)	88.11	31-91	91.03	73-133†	88.47	33-77
Straightforward QNQ mixed 1-3 word preambles (4.5-4.8)	79.22	113-189	90.88	71-129	87.64	53-95
Straightforward Q difficult NQ no preambles (4.33-4.36)	89.55	49-123	91.72	25-69	92.73	41-69
Simple Q Difficult NQ both with 2 word preambles (4.41-4.44)	89.03	53-107	91.63	19-27	93.04	15-23
Straightforward Q mixed S/D NQ no preambles (5.49-5.52)	89.37	59-101	93.14	27-47	93.63	49-73
Simulated clauses (4.49-4.52)	66.62	3-11	75.72	27-63	78.84	25-35

Table 5.64: A Comparison of the CAs (best performing trees) for the 3 different organisations of function words across a selection of the most demanding experiments performed in Chapter 4.

† The anomalously high range of tree sizes for the use of clustered function words with straightforward questions with 1 word preambles vs. straightforward and non-questions is the consequence of the improved CA. Selecting a parameter combination from the experiments that yields a CA close to the alpha sorted or grouped function words results in a consistent range of tree sizes. For example, a confidence level of 0.005 yields a CA of 88.33 and a range of tree sizes from 25-63.

5.11 Comparative analysis of experiments

The purpose of this section is to take an overview of the related groups of experiments and the evidence they provide about the general hypotheses. This involves:

- analysing the performance of clustered words vs. alphabetically sorted function words over the range of experiments
- analysing the performance of grouped words vs. alphabetically sorted function words over the range of experiments

5.11.1 The outcome of clustering function words

For each experiment reported in sections 5.7, the first hypothesis pair concerns improvement in terms of CA of clustered function words over alphabetically sorted function words. Table 5.65 summarises the improvements in the percentage CA (column 2) and identifies which are statistically significant (column 3).

Clustered Function Words Experiment	Improvement	Sig
Straightforward QNQ no preambles (4.1-4.4)	0.16	N
Straightforward QNQ 1 word preamble (4.9-4.12)	2.92	Y
Straightforward QNQ mixed 1-3 word preambles (4.5-4.8)	11.66	Y
Straightforward Q difficult NQ no preambles (4.33-4.36)	2.17	Y
Simple Q Difficult NQ both with 2 word preambles (4.41-4.44)	2.6	Y
Straightforward Q mixed S/D NQ no preambles (5.49-5.52)	3.77	Y
Simulated clauses (4.49-4.52)	9.1	Y

Table 5.65 Improvement in CA for Clustering compared with alphabetically sorting function words

The combination of straightforward questions with straightforward non- questions (neither with preambles) was the simplest. This was included as a check on whether changing the method of encoding function word features would have an impact on DTs which already had a high CA. The clustered form managed a very small improvement over alphabetic sorting, which was not statistically significant. All the other experiments show a statistically significant improvement in CA, with large increases for mixed length pre-ambles preceding simple questions (11.66) and for the simulated clauses (9.1).
Concerning the second hypothesis pair, the range of tree sizes is substantially reduced by moving from alphabetically sorted to clustered function words with two exceptions. For each of the exceptions (*straightforward questions with 1-word preambles vs.*

straightforward non-questions and *simulated clauses*), the increase in tree sizes can be explained by the possibility that deeper and more detailed structure has been discovered than was possible with alphabetically sorted function words and that the extra nodes are needed to encode this.

Consequently it is possible to conclude that the evidence from this series of experiments collectively provides quite strong evidence to support the two general forms of alternative (H_1) hypotheses:

- Clustering Function Words into functional groups improves classification accuracy over alphabetically sorted function words

- Clustering Function Words into functional groups does produce more compact trees than those from alphabetically sorted function words

5.11.2 The outcome of grouping function words

For each experiment reported in section 5.8, the first hypothesis pair concerns improvement in terms of CA of grouped function words over alphabetically sorted function words. Table 5.66 summarises the improvements in the percentage CA (column 2) and identifies which are statistically significant (column 3).

Grouped Function Words Experiment	improvement	Sig
Straightforward QNQ no preambles (4.1-4.4)	0.26	N
Straightforward QNQ 1 word preamble (4.9-4.12)	0.36	N
Straightforward QNQ mixed 1-3 word preambles (4.5-4.8)	8.42	Y
Straightforward Q difficult NQ no preambles (4.33-4.36)	3.18	Y
Simple Q Difficult NQ both with 2 word preambles (4.41-4.44)	4.01	Y
Straightforward Q mixed S/D NQ no preambles (5.49-5.52)	4.31	Y
Simulated clauses (4.49-4.52)	12.22	Y

Table 5.66 Improvement in CA for Grouping compared with alphabetically sorting function words

The first two combinations show an insignificant improvement over the alphabetically sorted function words approach. The experiment using mixed 1-3 word preambles with straightforward questions has a smaller improvement than the clustered function words (but still both significant and substantial). The final 4 experiments have CAs which consistently outperform the clustered function words.

Concerning the second hypothesis pair, the range of tree sizes is substantially reduced by moving from alphabetically sorted to clustered function words with two exceptions. In the case of *straightforward questions with 1-word preambles vs. straightforward non-questions* the smallest trees produced by alphabetically sorted function words are 2 nodes smaller than those produced using grouped function words, but the largest trees are markedly larger (91 nodes vs. 77). The other exception is *simulated clauses* and in both cases the increase in tree sizes can be explained by the possibility that deeper and more detailed structure has been discovered than was possible with alphabetically sorted function words and that the extra nodes are needed to encode this.

5.11.3 Comparison of clustered vs. grouped function words

The relative performance of the two approaches to improving feature extraction, through clustering or grouping the function words by function, is summarised in table 5.67. Note that negative entries in the table mean that grouping is performing less effectively than clustering. In the Sig column, a Y entry means the difference is statistically significant and N means it is not statistically significant.

Experiment	Improvement (grouped over clustered)	Sig
Straightforward QNQ no preambles (4.1-4.4)	0.1	N
Straightforward QNQ 1 word preamble (4.9-4.12)	-2.56	Y
Straightforward QNQ mixed 1-3 word preambles (4.5-4.8)	-3.24	Y
Straightforward Q difficult NQ no preambles (4.33-4.36)	1.01	N
Straightforward Q Difficult NQ both with 2 word preambles (4.41-4.44)	1.41	N
Straightforward Q mixed S/D NQ no preambles (5.49-5.52)	0.54	N
Simulated clauses (4.49-4.52)	3.12	Y

Table 5.67 Improvement in CA for Grouping compared with Clustering function words

In the simulated clauses experiment the grouping approach is clearly superior. However, when the questions are preceded by a 1-word preamble, or by a mix of 1 – 3 word preambles, the clustering approach is clearly superior. In two cases, where straightforward questions are combined with difficult non-questions which have either no preambles or 2-word preambles, the improvement is bordering on significance.

The evidence does not provide support for the rejection of the general null hypothesis:

> Grouping Function Words into functional groups does not improve classification accuracy over Clustered function words.

In favour of the general alternative hypothesis

> Grouping Function Words into functional groups improves classification accuracy over Clustered function words.

In 5 out of the 7 pairs of experiments, the grouped data DTs are smaller than the clustered data DTs (in 3 cases substantially smaller). The remaining 2 cases have substantially larger DTs for the grouped data.

These results provide limited evidence to reject the null hypothesis:

> Grouping Function Words into functional groups does not produce more compact trees than those produced by clustering function words into functional groups

In favour of the alternative hypothesis:

> Grouping Function Words into functional groups produces more compact trees than those produced by clustering function words into functional groups

5.12 Discussion

Having examined the results and hypotheses, it is appropriate to return to the original intuitions and explore possible causes of the experimental results before considering directions for future work.

5.12.1 Analysis of the original intuitions

Returning to the intuitions expressed at the start of the chapter, good evidence was found to support the first intuition. Tables 5.64 to 5.66 show that both clustering and grouping significantly improve on CA in all but the simplest cases. It is arguable that for the simplest case, the CA of the tree derived from alphabetically sorted function words has such a high CA (98.51) that there is little scope for improvement.

There is little evidence to support the second intuition, that grouping function words would lead to a further improvement in CA over clustering function words. Table 5.67 shows the balance is slightly in favour of the grouped function words in terms of the aggregate

improvement across the set of experiments but looking at individual results, only one shows significant improvement whilst two show significant deterioration in CA.
The third intuition is supported by 10 of the 14 experiments where the clustered or grouped function words trees are substantially smaller that those produced using alphabetically sorted function words. In the case of the straightforward questions and non-questions with 1 word preambles, the increased size can be explained, at least in part, as the price of achieving improved CA. In the case of simulated clauses, the argument that the alphabetically sorted function words have produced a trivial tree with a low CA (66.62) is quite convincing.
As for the final intuition, in 5 out of the 7 pairs of experiments comparing clustering with grouping, the grouped data DTs are smaller. In both of the contradictory cases, the increase in tree size is again the price of a small increase in CA.

5.12.2 Possible explanations for the relative performances of clustered vs. grouped features

5.12.2.1 Duplicated records

Grouping function words results in more abstract and generalised representation of the domain. Consequently, a number of questions or non-questions that have distinct representations under the alphabetically sorted or clustered coding become represented by a common record under grouped coding. Loss of these records may have penalised the grouped approach. If this is the explanation, then the grouped approach should achieve higher CA than the other approaches when encountering more diverse data in real-world applications.

5.12.2.2 Effect of restrictions on partitioning

The failure of grouping to improve over clustering could be explained by the fact that some of the function words have the ability to be a member of more than one group, as described in section 5.2. The clustered function words had limited discretion to adjust group boundaries during partitioning and this may have given them an advantage that compensated for the potentially greater generalising power of grouping.
Some insight may be gained from comparing a typical pair of DTs, using the two different techniques for the same dataset (straightforward questions vs. difficult non-questions, both without preambles). The trees are shown side-by-side in table 5.68.

Clustered	Grouped
w1 <= 215: q (220.0) w1 > 215 | w2 <= 187 | | w7 <= 257: n (556.0/51.0) | | w7 > 257: q (29.0/9.0) | w2 > 187 | | w2 <= 207: q (331.0/35.0) | | w2 > 207 | | | w7 <= 259: n (60.0/9.0) | | | w7 > 259: q (4.0)	w1 <= 188: q (211.0) w1 > 188 | w2 <= 141 | | w7 <= 254: n (557.0/51.0) | | w7 > 254: q (26.0/8.0) | w2 > 141 | | w2 <= 188 | | | w3 <= 52 | | | | w3 <= 0 | | | | | w9 <= 219: n (27.0/8.0) | | | | | w9 > 219: q (23.0/3.0) | | | | w3 > 0: q (216.0/4.0) | | | w3 > 52: n (29.0/12.0) | | w2 > 188: n (47.0/6.0)

Table 5.68 Comparison of pruned DTs for questions vs. difficult non-questions (both without preambles)

The following observations may be made:

- The splits in the clustered tree do not occur on the distinct boundaries between functional word groups used for the grouped approach.
- The top-level split (for w1) is on the word "must" (=215) and is ¾ of the way through the auxiliary verbs group.
- The first split on w2 is at the value for "without" (=187), this is at the upper boundary of the prepositions, immediately before the auxiliary verbs.
- The split on w7 is on the word "what" (=257), which is about 1/3 of the way through the interrogative introducers category.
- A further split on w2 occurs for the word "can" = 207. This is another split within the permission and ability auxiliary verbs, although there is some sense that the preceding auxiliary verbs deal with future events.
- The final split on w7 is on the word "when" (=259). This is about 2/3 of the way through the interrogative introducers.

So it appears that the clustered approach is adjusting the implicit boundaries of the functional grouping to optimise its performance over the dataset. On the other hand the grouped approach seems to be benefiting from the constraints imposed by conserving more information which allows a deeper tree to be constructed. The outcome is that, for the trees shown above, clustered data has produced a tree with a CA of 90.25 and grouped data has

produced one with a CA of 91.2. It could be observed that each approach has played to its own strengths and these have counterbalanced each other.

5.13 Chapter Conclusions

The experiments discussed in the chapter have revealed the potential for substantial and significant improvements in CA and size of DT classifiers for dialogue acts. The chapter produced a strong body of evidence to support this, derived from experiments covering a range of possible sentential contexts containing questions vs. non-questions. The most powerful classifier produced achieved a CA of 98.77 and several others had CAs in excess of 90. Analysis of the experimental outcomes revealed some genuinely intriguing questions regarding the construction of DTs which suggest the need for substantial and sustained future experimentation, on the basis that answering these questions will lead to further significant improvements in DA classifiers for DTs.

The success with this question/non-question challenge, provides some grounds for optimism in extending the function word classification technique to instruction / non-instruction and statement / non-statement classes, which are the essentials of human-machine dialogue systems.

Chapter 6 Discussion, Conclusions and Future Work

6.1 Discussion

The original motivation for the work in this thesis was to create a framework for the construction and evaluation of a new generation of Conversational Agents. As part of the work, certain key components and processes in the framework were identified and implemented. Chapter 1 showed that Conversational Agents have a long history in the field of computer science, beginning with the work of Alan Turing concerning a test that would prove acceptable evidence to humans that machines could think. The response from philosophers such as John Searle and Ned Block made an unwitting contribution to the development of Artificial Intelligence by sharpening the wits of the AI community, to quote Nietzsche

> "What does not kill me, makes me stronger" (Nietzsche, 1888).

In addition to explicit work, there has been substantial Conversational Agent related research since Turing in fields such as Dialogue Management or Question Answering. This begs the question, where are the developed Conversational Agents, assisting humans in our daily tasks? This thesis argues that the lack of a viable real-world application is largely due to issues of development cost, scalability and computational complexity. One potential tool in solving this problem is Short Text Semantic Similarity.

Semantic Similarity

Chapter 1 established the importance and acceptance of the concept of similarity as an experimental variable both historically and across a range of current scientific fields of application. It also explained how STSS has the potential to replace pattern matching (high labour costs in development and maintenance) and NLP (high computational overhead, chains of inaccurate processes) as a new technique for rapid agent development.

STSS-65: a Gold Standard benchmark dataset for evaluating STSS algorithms

When a new component, such as an STSS algorithm, is introduced into a complex system, such as a Conversational Agent, it is difficult to isolate the contribution of the particular component in an evaluation of the whole system. Yet the need for measurement in evaluation is crucial if we are to decide whether or not STSS is a worthy replacement for existing algorithms. As put by Lord Kelvin:

> "… when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind…" – Lord Kelvin (Thomson, 1889)

Therefore the work in chapter 2 concerned the production of an initial benchmark dataset for STSS algorithms and a systematic approach for using it to evaluate individual measures and compare a measure with others. During the course of the work, the best procedures for approaching ground truth similarity ratings were established and evidence was found that a robust underlying model of semantic similarity had been captured. This dataset has been published and analysed in a number of peer-reviewed destinations (Li et al., 2006, O'Shea et al., 2008, O'Shea, 2010). It is considered to be a major achievement as it has been adopted rapidly by other researchers in the field as a Gold Standard.

STSS-131: the new Gold Standard dataset

Chapter 3 was concerned with the production of a new STSS dataset with the aim of having the greatest coverage of features of the English Language possible within a restricted dataset. The novelty of the processes involved required the use of techniques from varied fields of study including neuropsychology, which provided an independent source of categories for word similarity.

Chapter 3 also used the SDT technique of discriminability which provided a powerful validation of the similarity ratings of the new sentence pairs created for the dataset. It was also used as an investigation of whether discriminablity could eliminate certain problems of questionnaire-based methods discussed in chapter 3. In practice the discriminability method was found to have disadvantages of its own and because of the lower productivity it was not considered a suitable replacement for the card sorting in the general measurement of STSS.

The new STSS Gold Standard benchmark dataset has not been published yet, but based on the reception of the initial Gold Standard dataset, it is expected that it will be adopted widely and will have longevity.

Chapter 3 ended with some intuitions about STSS algorithms that were gained through the process of constructing the benchmark datasets. The most important of these was the possibility of using dialogue act and propositional content as independent factors in an STSS model.

The Slim Function Word Classifier (SFWC) for dialogue acts

Chapter 4 explored the weaknesses of pattern matching and existing STSS measures in dealing with sentences containing particular dialogue acts. A proposal was made for

separation of dialogue acts from propositional content as part of a factor-based model (with a third factor, Valence) of STSS. Measuring STSS with this model requires accurate and computationally efficient classifiers for DAs and Valence. It was decided to conduct an experimental investigation of DA classifiers as the next stage of the work, specifically by developing a question vs. non-question DA classifier. This choice of questions was influenced by the value of question classification in the fields of Conversational Agents, Information Retrieval and Question Answering. Question recognition using pattern matching had also been a previous field of investigation by the MMU Intelligent Systems Group which provided a valuable background to the research.

A review of prior work on DA classification revealed the use of combinations of complex features. Some of these features were manually extracted and there was no prospect of automating the process. Others would be computationally inefficient (relying on chains of NLP processes). Consequently the emphasis in chapter 4 was on the development of a computationally efficient classifier, dubbed the Slim Function Word Classifier. This classifier was efficient because it used features based entirely on function words. In chapter 4 no assumptions were made about classes within the function words and they were simply organised in alphabetical order.

Optimisation of the SFWC

Chapter 5 compared two methods of optimising the function words by either grouping or clustering them according to function and grammatical properties. In both cases the intention was to assist the decision tree by reducing the number of partitions required to split the feature space at each level, preserving information for more effective splits at lower levels in the tree. Both techniques improved significantly on the simple alphabetical order of function words approach taken in chapter 4.

There was also a conjecture that the grouping approach would lead to further improvements over the clustered approach because it would produce more compact trees with better generalisation. The experiments did not find general evidence to support this, but in the case of the most difficult dataset, grouping produced a statistically significant improvement in classification accuracy over the clustered approach.

6.2 Conclusions

There are three aspects to developing a framework for applying short text semantic similarity in conversational agents:

- decomposing the conversational agent into a component architecture

- development of the novel components required by the architecture
- evaluation of the novel components before an agent is integrated

The component architecture

The work in this thesis assumes a component architecture which is fundamentally an evolutionary development of that presented in chapter 1 (figure 1.1), in which the dialogue agent uses STSS algorithms in place of pattern matching to decide which rule will be fired in response to a user utterance.

Evaluation of components (using existing STSS measures)

At the onset of the work, two STSS algorithms existed which were capable of being incorporated into a conversational agents, STASIS and LSA.

The evaluation objectives were achieved by developing two benchmark datasets STSS-65 (chapter 2) and STSS-131 (chapter 3); together with the necessary procedures to apply them so that results can be compared meaningfully for STSS algorithms developed by different research groups.

This work has been a success. The adoption of STSS-65 by the research community was reported in chapter 2. STSS-131 will be published shortly, but the experimental evaluation of STASIS and LSA in chapter 3 shows that broader representation of the English language has resulted in STSS-131 posing a far greater challenge to STSS algorithms.

The proposed new factor-based model of STSS

The factor-based model of STSS was derived from two sources. These were a substantial practical knowledge of the Dialogue Acts people use in communicating with Conversational Agents and an analysis of the behaviour of prior STSS measures using specially selected test cases. The practical knowledge was gained over several years, writing scripts for conversational agents in demanding applications such as debt advice. The resulting 3-factor model required development of new components, before it could be developed further.

Dialogue Act classification

The first step towards dialogue act classification was to build a series question / non-question DA classifier, using decision trees. The initial approach, taken in chapter 4, achieved a high classification accuracy (98.51%) on the most straightforward form of questions which was similar to that used in the TREC QA conference series datasets (discussed in chapter 1).

The performance of most of the classifiers in chapter 4 provided good "proof of concept" for function word based decision trees. These were optimised in chapter 5. Chapter 5 showed that a first attempt at optimisation resulted in classification accuracy of over 90% for all question forms apart from the most complex. It is difficult to draw any conclusions about the relative performances of particular techniques, because different studies use different features, different DA taxonomies and different corpora. One study (Levin et al., 2003) which used multi-classifiers over the whole range of dialogue acts reported a best combination of the k-Nearest-Neighbour and Bayesian classifiers as achieving 81.25% Classification Accuracy on English. Therefore the experiments in chapter 5 represent a significant achievement and are persuasive evidence that the necessary classifiers could be developed to support the proposed new factor-based model of STSS.

6.3 Future work

The work in this thesis addressed significant problems in the field of Short Text Semantic Similarity and has proposed novel solutions to problems of both validating STSS measures and devising new algorithms to overcome the limitations of existing work. However, the work itself has revealed that there is more to be done in both areas, evaluation and the development of the new model.

Conversational Agent evaluation

The bench datasets, STSS-65 and STSS-131 need to be expanded. The first issue is simply one of volume; more sentence pairs with similarity rating will support more extensive testing. It was a particular concern to see that some STSS researchers using STSS-65 had used the dataset in some way to train, tune or optimise their algorithms. This is to be discouraged as with a small dataset the improvements can come from optimisation for the dataset rather than generalisation to the whole population (of ST pairs in this case). However, if the dataset could be expanded to several hundred pairs it would be feasible to use cross-validation to train classifiers.

The other issue is that despite the efforts made to promote DA diversity; most of the sentences collected for STSS-131 were assertion DAs. Further data collection experiments are to promote DA diversity (this would require careful design of some additional constraints).

It should also be possible to expand the number of stimulus words. This could be done by:

- use of additional Battig & Montague categories for nouns
- use of finer grained categories for verbs (re-visiting RRG and Case Grammar)

Factor-based model of STSS

Expansion of the dataset is also an important element of the development of the factor-based model of STSS. For the model to be effective, the functions that weight the contributions of the three factors must be identified. This involves:

- the creation of sentence pairs that vary factors
- the collection of similarity ratings following the processes described in this thesis
- the statistical analysis of the results to extract the weighting functions

Varying the factors will involve varying one of the factors whilst holding two constant and varying two of the functions whilst holding one constant. This will be necessary to get a full picture of potential interactions between the factors. Consequently a large number of permutations can be generated and the design of the experiments involved will be quite complex.

Development of components for the factor-based STSS model

Chapter 5 has already shown the potential for optimising the SFWC for questions. The difficulties discussed in chapter 5 regarding the clustering and grouping approaches can be tackled in a number of ways, in particular:

- The clustered approach could be optimised using a genetic algorithm
- The grouped approach could be optimised using Fuzzy Logic

Optimising the clustered approach would involve creating a population of chromosomes which were representations of the function word table. In each chromosome the order of function words in the table would be randomised. Through processes of crossover and mutation (with classification accuracy as the cost function), the population would be evolved to discover better clusters of function words. This work would require initial research on GA techniques where each allele must be unique.

Optimising the grouped approach would involve creating fuzzy sets based on the functional / grammatical classes used in chapter 5 and assigning a degree of set membership to each function word for each set. This should overcome the information lost in the crisp approach where a word which belongs to several classes is forced into one only.

Finally it is possible that alternative tokenising approaches, such as using counts of content word strings (instead of strings of zeros) could promote generalisation through more abstract representation.

Development of classifiers for different DAs

Experience in developing pattern matching Conversational Agents suggests that the most important distinction for STSS measurement will be between statements, questions and instructions. Therefore a minimum of two additional classifiers are required:

- A classifier for statements / non-statements
- A classifier for instructions / non-instructions

Future work will require a replication of the processes in chapters 4 and 5 for both these classifier types. The final stage of development is likely to be a multi-classifier combining the outputs from individual DA classifiers to reach a decision. This multi-classifier will also open the way to producing a classifier for complex sentences containing more than one DA.

The "Take-Home" message

What is the take-home message from this thesis? The CA of the future could be so much more than the advisory and selling applications focused on today. They could become the crucial empowering technology of the future by allowing us to instruct agents that act on our behalf in complex multi-agent negotiating systems.

For example, the smart grid addresses problems associated with the traditional supply of electricity where inefficient and expensive generating plant must be activated in response to transient increases in demand. It incorporates micro storage in which consumers store an even generate their own electricity. This allows the grid to sell electricity at varying price to consumers at different times and even buy it back from them at times of peak demand. Managing the smart grid involves numbers of agents negotiating the transfers of power. Imagine the following conversation:

<user> I want you to be careful about managing my electricity tonight.
<agent> OK, in what way?
<user> I'm driving to the airport and I want my car fully charged.
<agent> Swansea Hydro are buying regularly during the early evening, what time to you leave for the airport?
<user> 4:00 am.
<agent > which airport are you going to?
<user> Manchester international

> <agent> Your car battery half full and your home battery is three-quarters full. I can sell up to half of what's in your home battery to Swansea Hydro and still cover your needs tomorrow by topping up at the cheaper rate.
> <user> OK – do it if they ask.

And so some Conversationsl Agents become Intentional Agents expressing our beliefs, desires and intentions in the real world. As the average human does not have the capability of instructing such agents using formal specification languages or temporal logic. The conversational aspect becomes crucial. Existing technologies (Pattern Matching and NLP) do not have the appropriate characteristics for producing highly accurate, computationally efficient and above all maintainable CAs.

Therefore the message to all CA developers is to embrace new technologies and algorithms such as semantic similarity measurement and DA classification. To do this, they will be supported by elements of the framework described in this thesis.

References

ACHANANUPARP, P., HU, X. & SHEN, X. (2008) The Evaluation of Sentence Similarity Measures. *Lecture Notes in Computer Science,* 5182**,** 305–316.

AGIRRE, E. & RIGAU, G. (1995) A Proposal for Word Sense Disambiguation using Conceptual Distance. *1st Intl. Conf. on recent Advances in NLP.* . Bulgaria.

AIMLESS, D. Pandorabots - A Multilingual Chatbot Hosting Service.

ALLEN, J. & CORE, M. (1997) Draft of DAMSL: Dialog Act Markup in Several Layers. University of Rochester, Rochester, USA. .

ANDERNACH, T. (1996) A Machine Learning Approach to the Classification of Dialogue Utterances.

ANDERNACH, T., POEL, M. & SALOMONS, E. (1997) Finding Classes of Dialogue Utterances with Kohonen Networks

ECML/MLnetWorkshop on Empirical Learning of Natural Language Processing Tasks. Prague, Czech Republic.

ANDERSEN, V., ANDERSEN, H. H. K., DEGEMMIS, M., LICCHELLI, O., LOPS, P. & ZAMBETTA, F. (2003) A methodological approach for designing and evaluating intelligent applications for digital collections. *Applied Artificial Intelligence,* 17**,** 745-771.

ANJOMSHOAA, A., KARIM, S., SHAYEGANFAR, F. & TJOA, A. M. (2006) Exploitation of Semantic Web Technology in ERP Systems *Research and Practical Issues of Enterprise Information Systems* Boston Springer

AQA (2010) AQA Languages.

ARONOFF , J. M., GONNERMANB, L. M., ALMORC, A., ARUNACHALAM, S., KEMPLER, D. & ANDERSEN, E. S. (2006) Information content versus relational knowledge: Semantic deficits in patients with Alzheimer's disease. *Neuropsychologia,* 44 21–35.

AUGELLO, A., VASSALLO, G., GAGLIO, S. & PILATO, G. (2008) Sentence Induced Transformations in "Conceptual" Spaces. *International Conference on Semantic Computing, 2008 IEEE.* Santa Clara, CA,, IEEE.

AUSTIN, J. L. (1975) *How to do things with Words: The William James Lectures delivered at Harvard University in 1955*, Harvard University Press.

BABU, S., SCHMUGGE, S., BARNES, T. & HODGES, L. F. (2006) "What Would You Like to Talk About?" An Evaluation of Social Conversations with a Virtual Receptionist. *Lecture Notes in Computer Science* 4133/2006

BALLARDINI, N., YAMASHITA, J. A. & WALLACE, W. P. (2007) Presentation duration and false recall for semantic and phonological associates. *Consciousness and Cognition,* 17**,** 64-71.

BANGALORE, S. & STENT, A. J. (2009) Incremental parsing models for dialog task structure. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.* Athens, Greece

BATTIG, W. F. (1979) Citation Classic: Battig W F & Montague W E. Category norms for verbal items in 56 categories: a replication and extension of the Connecticut category norms. J. Exp. Psychol. Monograph Suppl. 80: No. 3, Part 2, 1969. IN GARFIELD, E. (Ed.).

BATTIG, W. F. & MONTAGUE, W. E. (1969) Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. . *Journal of Experimental Psychology Monographs* 80**,** 1–46.

BBC (2009) Tories criticise flu advice line BBC.

BÉDARD, A. & PARENT, A. (2004) Evidence of newly generated neurons in the human olfactory bulb. *Brain Res Dev Brain Res.* , 151**,** 159-68.

BEDNY, M. & THOMPSON-SCHILL, S. L. (2006) Neuroanatomically separable effects of imageability and grammatical class during single-word comprehension. *Brain and Language,* 98 127–139.

BENET-MARTINEZ, V. & WALLER, N. G. (2002) From adorable to worthless: Implicit and self-report structure of highly evaluative personality descriptors. *European journal of personality* 16**,** 1-41.

BICKMORE, T. & CASSELL, J. (2000) 'How about this weather?' Social Dialog with Embodied Conversational Agents. *The American Association for Artificial Intelligence (AAAI) Fall Symposium on "Narrative Intelligence".* Cape Cod, MA. .

BICKMORE, T. & GIORGINO, T. (2006) Health dialog systems for patients and consumers. *J Biomed Inform,* 39**,** 556-571.

BILMES, G. J. A. J. (2005) Dialog act tagging using graphical models *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Philadelphia, USA

BLALOCK, H. M. (1979) *Social Statistics* McGraw-Hill Inc.

BLOCK, N. (1981) Psychologism and behaviourism. *The Philosophical Review,* LXXXX**,** 5-43.

BOHUS, D. & HORVITZ, E. (2009a) Learning to Predict Engagement with a Spoken Dialog System in Open-World Settings. IN HEALEY, P., PIERACCINI, R., BYRON, D., YOUNG, S. & PURVER, M. (Eds.) *The SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* London, UK.

BOHUS, D. & HORVITZ, E. (2009b) Models for Multiparty Engagement in Open-World Dialog

IN HEALEY, P., PIERACCINI, R., BYRON, D., YOUNG, S. & PURVER, M. (Eds.) *The SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* London, UK.

BOLLACKER, K. D., LAWRENCE, S. & GILES, C. L. (1998) CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications. *2nd International ACM Conference on Autonomous Agents.* ACM Press.

BONDARIONOK, A., BOBKOV, A., SUDANOVA, L., MAZUR, P. & SAMUSEVA, T. (2007) Intellexer Question Answering. *The Sixteenth Text REtrieval Conference.* NIST.

BOUWMAN, G. & HULSTIJN, J. (1998) Dialog Strategy Redesign with Reliabilty Measures. *1st Int. Conf. on Language Resources and Evaluation.* Granada, Spain.

BOUWMAN, G., STURM, J. & BOVES, L. (1999) Incorporating confidence measures in the Dutch train timetable information system developed in the ARISE project. *ICASSP '99.*

BOYE, J., WIREN, M. & GUSTAFSON, J. (2004) Contextual reasoning in multimodal dialogue systems: two case studies. . *The 8th Workshop on the Semantics and Pragmatics of Dialogue Catalogue'04* Barcelona.

BREEDIN, S. D., SAFFRAN, E. M. & COSLETT, H. B. (1994) Reversal of the concreteness effect in a patient with semantic dementia *Cognitive Neuropsychology,* 11**,** 617 - 660.

BREIMAN, L., JFRIEDMAN, J. H., OLSHEN, R. A. & STONE, C. J. (1984) *Classification and Regression Trees.*, Wadsworth.

BRINGSJORD, S., BELLO, P. & FERRUCCI, D. (2001) Creativity, the Turing Test and the (Better) Lovelace Test. *Minds and Machines,* 11**,** 3-27.

BUI, T. H., POEL, M., NIJHOLT, A. & ZWIERS, J. (2007) A tractable DDN-POMDP approach to affective dialogue modeling for general probabilistic frame-based dialogue systems *International Joint Conference on AI, IJCAI07* India, 2007.

BUSHELL, C. M. & MARTIN, A. (1997) Automatic semantic priming of nouns and verbs in patients with Alzheimer's disease. *Neuropsychologia,* 35**,** 1059 1067.

CAPITANI, E., LAIACONA, M., MAHON, B. Z. & CARAMAZZAZ, A. (2003) What are the facts of Semantic Category-Specific Deficits? A critical review of clinical evidence. *Cognitive Neuropsychology,* 20**,** 213-261.

CAPUANO, N., GAETA, M., ORCIUOLI, F. & RITROVATO, P. (2009) Ninth IEEE International Conference on Advanced Learning Technologies. *On-Demand Construction of Personalized Learning Experiences Using Semantic Web and Web 2.0 Techniques.* IEEE Computer Society Washington, DC, USA

CARAMAZZA, A. (1998) The interpretation of semantic category-specific deficits: what do they reveal about the organisation of conceptual knowledge in the brain? *Neurocase,* 4.

CARAMAZZA, A., HILLIS, A. E., RAPP, B. C. & ROMANI, C. (1990) The multiple semantics hypothesis: Multiple confusions? *Cognitive Neuropsychology,* 7**,** 161-189.

CARAMAZZA, A. & SHELTON, J. R. (1998) Domain-Specific Knowledge Systems in the Brain: The Animate-Inanimate Distinction. *Journal of Cognitive Neuroscience* 10**,** 1–34.

CARDIE, C., VINCENT NG, V., PIERCE, D. & BUCKLEY, C. (2000) Examining the Role of Statistical and Linguistic Knowledge Sources in a General-Knowledge Question-Answering System. *The Sixth Applied Natural Language Processing Conference (ANLP-2000).* Association for Computational Linguistics / Morgan Kaufmann

CARTER, C. & MCCARTHY, M. (2006) *Cambridge Grammar of English*, Cambridge University Press.

CASSELL, J. & BICKMORE, T. (2003) Negotiated Collusion: Modeling Social language and its Relationship Effects in Social Agents. *User Modeling and User-Adapted Interaction* 13**,** 89-132.

CASSELL, J., BICKMORE, T., VILHJALMSSON, H. & YAN, H. (2001) More Than Just a Pretty Face: Conversational Protocols and the Affordances of Embodiment. *Knowledge-Based Systems,* 14**,** 55-64.

CASSELL, J., SULLIVAN, J., PREVOST, S. & CHURCHILL, E. (2000) *Embodied Conversational Agents*.

CASSELL, J. & VILHJÁLMSSON, H. (1999) Fully Embodied Conversational Avatars: Making Communicative Behaviors Autonomous. *Autonomous Agents and Multi-Agent Systems,* 2**,** 45-64.

CHAFE , W. L. (1970) *Meaning and the structure of language*, University of Chicago Press.

CHAO, L. L., JAMES V. HAXBY, J. V. & MARTIN, A. (1999) Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *nature neuroscience* 2

CHARLES, W. G. (2000) Contextual Correlates of Meaning. *Applied Psycholinguistics* 21**,** 505-524.

CHEN, L.-H., LAI, Y.-C. & LIAO, H.-Y. M. (2008) Movie scene segmentation using background information. *Pattern Recognition* 41**,** 056 – 1065.

CHIEN, S. & IMMORLICA, N. (2005) Semantic Similarity Between Search Engine Queries Using Temporal Correlation. *International World Wide Web Conference (IW3C2).* Chiba, Japan.

CHRISTENSEN, C. R. & GARVIN, D. A. (1992) *Education for Judgment: The Artistry of Discussion Leadership*, Harvard Business School Press.

CHWILLA, D. J., KOLK, H. H. J. & VISSERS, C. T. W. M. (2007) Immediate integration of novel meanings: N400 support for an embodied view of language comprehension. *BRAIN RESEARCH,* 1183**,** 109–123.

CLARK, A. & POPESCU-BELIS, A. (2004) Multi-level Dialogue Act Tags. *5th SIGDIAL Workshop on Discourse and Dialog, SIGDIAL '04* Cambridge, MA.

CONEY, B. AskOxford: How many words are there in the English language? , Oxford Dictionaries

CONVERSIVE, I. Verbots - Open Source Natural Language Chatterbots.

COOK, W. (1979) *Case Grammar: development of the Matrix Model (1979-1978)*, Georgetown University Press

COSLETT, H. B. & LIE, E. (2008) Simultanagnosia: Effects of semantic category and repetition blindness. *Neuropsychologia* 46**,** 1853–1863.

COXHEAD, A. (2000) A new academic word list. *TESOL Quarterly,* 34**,** 213-238.

CREE, G. S. & MCRAE, K. (2003) Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). . *Journal of Experimental Psychology: General,* 132, 163–201.

CROCKETT, K., BANDAR, Z., O'SHEA, J. & MCLEAN, D. (2009) Bullying and Debt: Developing Novel Applications of Dialogue Systems. *Knowledge and Reasoning in Practical Dialogue Systems (IJCAI).* Pasadena, CA, IJCAI.

CROSS, K., SMITH, E. E. & GROSSMAN, M. (2008) Knowledge of natural kinds in semantic dementia and Alzheimer's disease. *Brain and Language,* 105**,** 32–40.

DAMASIO, H., TRANEL, D., GRABOWSKI, T. J., ADOLPHS, R. & DAMASIO, A. R. (2004) Neural systems behind word and concept retrieval. *Cognition,* 92**,** 179–229.

DAMIANO, R., GENA, C., LOMBARDO, V., NUNNARI, F. & PIZZO, A. (2008) A stroll with Carletto: adaptation in drama-based tours with virtual characters. *User Modeling and User-Adapted Interaction,* 18**,** 417-453.

DANG, H. T., DIANE KELLY & LIN, J. (2007) Overview of the TREC 2007 Question Answering Track. *The Sixteenth Text REtrieval Conference.* NIST.

DAVIDSON, G. (Ed.) (2004) *Roget's Thesaurus of English Words and Phrases*, Penguin.

DE BONI, M. & MANANDHAR, S. (2003) The Use of Sentence Similarity as a Semantic Relevance Metric for Question Answering. *The AAAI Symposium on New Directions in Question Answering.* Stanford, AAAI Press

DE ZUBICARAY, G. I., MCMAHON, K. L., EASTBURN, M. M. & PRINGLE, A. J. (2008) Negative priming in naming of categorically related objects: An fMRI study. *cortex* 44 881–889.

DEERWESTER, S., DUMAIS, S., FURNAS, G. W., HARSHMAN, R., LANDAUER, T., LOCHBAUM, K. & STREETER, L. (1989) Computer information retrieval using Latent Semantic Structure. IN OFFICE, U. S. P. (Ed.) United States of America, Bell Communications Research Inc.

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. & HARSHMAN, R. (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science,* 41**,** 391-407.

DEHAENE, S. & NACCACHE, L. (2001) Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition,* 79 1-37.

DENNIS, N. A., KIMB, H. & CABEZA, R. (2007) Effects of aging on true and false memory formation: An fMRI study. *Neuropsychologia* 45**,** 3157–3166.

DESCARTES, R. (Ed.) (1988) *The Philosophical Writings Of Descartes,* , Cambridge, Cambridge University Press.

DESMONTILS, E. & JACQUIN, C. (2002) Indexing a Web Site with a Terminology Oriented Ontology. *et02 The Emerging Semantic Web.* IOS Press.

DEVAULT, D., SAGAE, K. & TRAUM, D. (2009) Can I finish? Learning when to respond to incremental interpretation results in interactive dialogue IN HEALEY, P., PIERACCINI, R., BYRON, D., YOUNG, S. & PURVER, M. (Eds.) *The SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* London, UK.

DEVLIN, J. T., RUSSELL, R. P., DAVIS, M. H., PRICE, C. J., MOSS, H. E., FADILI, M. J. & TYLER, L. K. (2002) Is there an anatomical basis for category-specificity? Semantic memory studies in PET and fMRI *Neuropsychologia,* 40**,** 54-75.

DHYANI, D., NG, W. K. & BHOWMICK, S. S. (2002) A Survey of Web Metrics. *ACM Computing Surveys,* 34**,** 469-503.

DIEN, J. (2008) Looking both ways through time: The Janus model of lateralized cognition. *Brain and Cognition.*

DIMMICK, D. (2000) Introduction (to PRISE). NIST.

DIXON, R. M. W. (1991) *A new approach to English Grammar, on Semantic Principles*, Oxford University Press: Clarendon Paperbacks.

DIXON, R. M. W. & AIKHENVALD, A. (2004) *Adjective Classes: A Cross-Linguistic Typology.,* Oxford University Press

DOHSAKA, K., ASAI, R., HIGASHINAKA, R., MINAMI, Y. & MAEDA, E. (2009) Effects of Conversational Agents on Human Communication in Thought-Evoking Multi-Party Dialogues. IN HEALEY, P., PIERACCINI, R., BYRON, D., YOUNG, S. & PURVER, M. (Eds.) *The SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* London, UK.

DRAKOS, N. (1998) Verb Semantic Classes. EAGLES Central Secretariat.

EGAN, J. P. (1975) *Signal detection theory and ROC analysis.,* London Academic Press.

ERION, G. J. (2001) The Cartesian Test for Automatism. *Mind and Machines,* 11**,** 29-39.

ERKAN, G. & RADEV, D. R. (2004) LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Articial Intelligence Research,* 22**,** 457-479.

EROZEL, G., CICEKLI, N. K. & CICEKLI, I. (2008) Natural language querying for video databases. *Information Sciences* 178 2534–2552.

FAATZ, A. & STEINMETZ, R. (2002) Ontology Enrichment with Texts from the WWW. *ECML-Semantic Web Mining.* Helsinki Finland.

FARAH, M. J. & MCCLELLAND, J. L. (1991) A computational model of semantic memory impairment: Modality specifi□city and emergent category specificity. . *Journal of Experimental Psychology: General,* 120**,** 339-357.

FARMER, R. (1996) *Teach Yourself Russian*, Hodder & Stoughton.

FARQUHAR, A., FIKES, R. & RICE, J. (1997) The Ontolingua Server: a Tool for Collaborative Ontology Construction. *Journal of Human-Computer Studies, ,* 46:**,** 707-728

FENG, J., ZHOU, Y. & MARTIN, T. (2008) Sentence Similarity based on Relevance. *IPMU 2008.* Torremolinos.

FENTON, N. & PFLEEGER, S. (1998) *Software Metrics: A Rigorous and Practical Approach. ,* PWS

FERNANDEZ, R., FRAMPTON, M., EHLEN, P., PURVER, M. & S., P. (2008) Modelling and Detecting Decisions in Multi-party Dialogue. *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue.* Columbus, Ohio, USA.

FERNANDEZ, R. & PICARD, R. W. (2002) Dialog Act Classification from Prosodic Features Using Support Vector Machines. *Speech Prosody.*

FERRI, F., GRIFONI, P. & PAOLOZZI, S. (2007) An Approach to Multimodal Input Interpretation in Human-Computer Interaction. *The Nineteenth International Conference on Software Engineering Knowledge Engineering (SEKE'2007),.* Boston, Massachusetts, USA, .

FINKELSTEIN, L., GABRILOVICH, E., MATIAS, Y., RIVLIN, E., SOLAN, Z., WOLFMAN, G. & RUPPIN, E. (2001) Placing Search in Context: The Concept Revisited. *The Tenth International World Wide Web Conference (WWW10)*

FLYNN, R. (2002a) Abstract Nouns - Glossary Definition - UsingEnglish.com.

FLYNN, R. (2002b) Question types - Glossary Definition - UsingEnglish.com.

FORDE, E. M. E., FRANCIS, D., RIDDOCH, M. J., RUMIATI, R. I. & HUMPHREYS, G. W. (1997) On the Links between Visual Knowledge and Naming: a Single Case Study of a Patient with a Category-specific Impairment for Living Things. *COGNITIVE NEUROPSYCHOLOGY,* 14, 403–458.

FORDE, E. M. E. & HUMPHREYS, G. W. (2002) *Category specificity in brain and mind.,* East Sussex, UK, Psychology Press.

FORMICA, A. (2008) Concept similarity in Formal Concept Analysis: An information content approach. *Knowledge-Based Systems,* 21**,** 80-87.

FRANCIS, W. N. & KUCERA, H. (1982) *Frequency Analysis of English Usage: Lexicon and Grammar,* Boston, Houghton Mifflin

FRENCH, R. M. (2001) Why co-occurrence information alone is not sufficient to answer subcognitive questions. *J. Expt. Theor. Artificial Intelligence* 13**,** 421-429.

FUNNELL, E. & SHERIDAN, J. S. (1992) Categories of knowledge? Unfamiliar aspects of living and nonliving things. . *Cognitive Neuropsychology,* 9**,** 135-153.

GACITUA-DECAR, V. & PAHL, C. Automatic Business Process Pattern Matching for Enterprise Services Design. *2009 World Conference on Services - II.* Bangalore, India

GAINOTTI, G. & SILVERI, M. C. (1996) Cognitive and Anatomical Locus of Lesion in a Patient with a Category-pecic Semantic Impairment for Living Beings. *COGNITIVE NEUROPSYCHOLOGY,* 13**,** 357-389.

GLASS, J., WEINSTEIN, E., CYPHERS, S., POLIFRONI, J., G., C. & NAKANO, N. (2004) A Framework for Developing Conversational User Interfaces. *Fourth International Conference on Computer-Aided Design of User Interfaces.* Funchal, Isle of Madeira, Portugal.

GLEITMAN, L. R., GLEITMAN, H., MILLER, C. & OSTRIN, R. (1996) Similar, and similar concepts. *Cognition,* 58**,** 321-376.

GOH, O. S., DEPICKERE, A., FUNG, C. C. & W., W. K. (2007) A Multilevel Natural Language Query Approach for Conversational Agent Systems. *IAENG International Journal of Computer Science,* 33.

GOLDBERG, D. E. (1989) *Genetic algorithms in search, optimization, and machine learning* Addison-Wesley.

GORIN, A. L., RICCARDI, G. & WRIGHT, J. H. (1997) How may I help you? *Speech Communication,* 23**,** 113-127.

GOSSELIN, P. H., CORD, M. & PHILIPP-FOLIGUET, S. (2008) Combining visual dictionary, kernel-based similarity and learning strategy for image category retrieval. *Computer Vision and Image Understanding* 110 403–417.

GRAESSER, A. C., CHIPMAN, P., HAYNES, B. C. & OLNEY, A. (2005) AutoTutor: An Intelligent Tutoring System With Mixed Initiative Dialogue. *IEEE Transactions on Education,* 48**,** 612-618.

GREEN, M. (2007) Speech Acts (Stanford Encyclopaedia of Philosophy). *Stanford Encyclopaedia of Philosophy.* Stanford, Stanford University.

GREEN, S. J. (1997) Building hypertext links in newspaper articles using semantic similarity. *Third Workshop on Applications of Natural Language to Information Systems (NLDB '97).* Vancouver, Canada.

GRIMNES, G. A., EDWARDS, P. & PREECE, A. (2008) Instance Based Clustering of Semantic Web Resources. *Lecture Notes in Computer Science,* 5021**,** 303-317.

GRIOL, D., CALLEJAS, Z. & LOPEZ-COZAR, R. (2009) A Comparison between Dialog Corpora Acquired with Real and Simulated Users. IN HEALEY, P., PIERACCINI, R., BYRON, D., YOUNG, S. & PURVER, M. (Eds.) *The SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* London, UK.

GUEST, E. (2008) Parsing for Role and Reference Grammar. IN VAN VALIN, R. D. (Ed.) *Investigations of the Syntax-semantics-pragmatics Interface (Studies in Language Companion).* John Benjamins Publishing Co.

GUREVYCH, I. & STRUBE, M. (2004) Semantic Similarity Applied to Spoken Dialogue Summarization. *20th International Conference on Computational Linguistics.* Geneva, Switzerland.

HACIOGLU, K. & WARD, W. (2003) Question Classification with Support Vector Machines and Error Correcting Codes. *HLT-NAACL 2003, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* Edmonton, Canada.

HALKIDI, M., NGUYEN, B., VARLAMIS, I. & VAZIRGIANNIS, M. (2003) Thesus: Organising Web Document Collections based on Semantics and Clustering. *Journal on Very Large Databases Special Edition on the Semantic Web*

HAN, L., SUN, L., CHEN, G. & XIE, L. (2006) ADSS: An approach to determining semantic similarity. *Advances in Engineering Software* 37**,** 129–132.

HANKS, P., MACLEOD, W. & L., U. (1986) *Collins Dictionary of the English Language*, William Collins Sons & Co.

HARTFIELD , K. N. & CONTURE, E., G. (2006) Effects of perceptual and conceptual similarity in lexical priming of young children who stutter: Preliminary findings. *Journal of Fluency Disorders,* 31 303–324.

HASSAN, S., CSOMAI, A., BANEA, C., SINHA, R. & MIHALCEA, R. (2007) UNT: SubFinder: Combining Knowledge Sources for Automatic Lexical Substitution. *SemEval, 4th International Workshop on Semantic Evaluations.*

HASSEBRAUCK, M. & BUHL, T. (1996) Three-dimensional love. *The Journal of Social Psychology,* 136.

HATZIVASSILOGLOU, V., KLAVANS, J. L. & ESKIN, E. (2001) Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning, *NAACL.*

HAWKIN, T. (1995) *Letts Study Guide Key stage 3 German* Letts.

HE, Y. & YOUNG, S. (2005) Semantic processing using the Hidden Vector State model. *Computer Speech & Language,* 19**,** 85-106

HERZ, R. S., ELIASSEN, J., BELAND, S. & SOUZA, T. (2004) Neuroimaging evidence for the emotional potency of odor-evoked memory. *Neuropsychologia* 42 371-378.

HEYLIGHEN, F. (2001) Mining Associative Meanings from the Web: from word disambiguation to the global brain. IN TEMMERMAN, R. & LUTJEHARMS, M. (Eds.) *Proceedings of the International Colloquium: Trends in Special Language & Language Technology.* Standaard Editions, Antwerpen.

HIRSCHBERG, J. & LITMAN, D. (1993) Empirical studies on the disambiguation of cue phrases. *Computational Linguistics,* 19**,** 501-530.

HIRSCHMAN, L. & GAIZAUSKAS, R. (2001) Natural Language Question Answering: The View from Here *Journal of Natural Language Engineering* 7**,** 275-300.

HSIEH, S. & HSU, C.-C. (2008) Retrieval of images by spatial and object similarities. *Information Processing and Management* 44 1214–1233.

HU, W., QU, Y. & CHENG, G. (2008) Matching large ontologies: A divide-and-conquer approach. *Data & Knowledge Engineering,* 67**,** 140-160.

HUANG, J.-J., CHANGT, S.-T. & HU, S.-Y. (2008) *Consumer Communications and Networking Conference, 2008. CCNC 2008. 5th IEEE.* Las Vegas, NV, .

HUETTIG, F., QUINLAN, P. T., MCDONALD, S. A. & ALTMANN, G. T. M. (2006) Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychologica* 121 65–80.

HUGHES, B., BIRD, S., LEE, H. & KLEIN, E. (2004) Experiments with Data-Intensive NLP on a Computational Grid. *The International Workshop on Human Language Technology* University of Hong Kong.

HUNDSNURSCHER, F. & SPLETT, J. (1982.) *Semantik der Adjektive im Deutschen:Analyse der semantischen Relationen* Westdeutscher Verlag Opladen.

HUNT, E. B., MARIN, J. & STONE, P. J. (1966) *Experiments in Induction* New York, Academic Press.

HUNT, M. (2001) *Teach Yourself English*, Hodder & Stoughton.

IARPA IARPA - What Is IARPA?

IDE, N. C. & VERONIS, J. (1998) Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics,* 24.

INABA, M., NOMURA, M. & OHIRA, H. (2005) Neural evidence of effects of emotional valence on word recognition. *International Journal of Psychophysiology,* 57**,** 165-173

INKPEN, D. (2007) Semantic Similarity Knowledge and its Applications. *Studia Universitatis Babes-Bolyai Informatica,* LII.

IRS (2009) Frequently Asked Tax Questions and Answers.

ISLAM, A. & INKPEN, D. (2008) Semantic Text Similarity using Corpus-Based Word Similarity and String Similarity. *ACM Transactions on Knowledge Discovery from Data,* 2**,** 1–25.

IVANOVIC, E. (2005) Dialogue Act Tagging for Instant Messaging Chat Sessions. *The ACL Student Workshop 2005.* Ann Arbor, Michigan, Association for Computational Linguistics.

IYER, N., JAYANTI, S., LOU, K., KALYANARAMAN, Y. & RAMANI, K. (2005) Shape-based searching for product lifecycle applications. *Computer-Aided Design* 37 1435–1446.

JACKENDOFF, R. (1972) *Semantic Interpretation in Generative Grammar*, MIT press.

JACKENDOFF, R. (1983) *Semantics and Cognition,* Cambridge, Massachusetts MIT Press.

JACKENDOFF, R. (1990) *Semantic Structures,* Cambridge, Massachusetts, MIT press.

JACKSON, P. & MOULINIER, I. (2007) *Natural Language Processing for Online Applications,* Amsterdam, The Netherlands, John Benjamins Publishing Company.

JEON, J., CROFT, W. B. & LEE, J. (2005) Finding Similar Questions in Large Question and Answer Archives. *The ACM Fourteenth Conference on Information and Knowledge Management (CIKM 2005).*

JEONG, B., LEE, D., CHO, H. & LEE, J. (2008) A novel method for measuring semantic similarity for XML schema matching. *Expert Systems with Applications,* 34**,** 1651-1658.

JEONG, B., LEE, D., LEE, J. & CHO, H. (2009) Support for seamless data exchanges between web services through information mapping analysis using kernel methods. *Expert Systems with Applications,* 36 358-365

JIANG, J. J. & CONRATH, D. W. (1997) Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *International Conference Research on Computational Linguistics (ROCLING X),* Taiwan.

JOKINEN, K., HURTIG, T., K., H., KANTO, K., KAIPAINEN, M. & KERMINEN, A. (2001) Self-Organizing Dialogue Management IN ISAHARA, H. & MA, Q. (Eds.) *The 2nd Workshop on Natural Language Processing and Neural Networks, NLPRS2001.* Tokyo, Japan.

JONES, M. N., KINTSCH, W. & MEWHORT, D. J. K. (2006) High-dimensional semantic space accounts of priming. *Journal of Memory and Language,* 55**,** 534–552.

JOUN, S., YI, E., RYU, C. & KIM, H. (2003) A Computation of Fingerprint Similarity Measures Based on Bayesian Probability Modeling *Lecture Notes in Computer Science* 2756**,** 512-520

JOVANOVIC, N. & OP DEN AKKER, R. (2004) Towards automatic addressee identification in multi-party dialogues. *The 5th SIGdial Workshop on Discourse and Dialogue.*

JOVANOVIC, N., OP DEN AKKER, R. & NIJHOLT, A. (2005) A corpus for studying addressing behavior in multi-party dialogues. I. *6th SIGdial Workshop on Discourse and Dialogue.* Lisbon, Portugal

JUOLA, P. (1996) LINGUIST List 7.583: Am. English word frequency lists. IN SEELY, T. D. (Ed.) *LINGUIST List*

KAROV, Y. & EDELMAN, S. (1998) Similarity-based Word Sense Disambiguation. *Computational Linguistics,* 24**,** 41 - 59

KATZ, B., BORCHARDT, G. & FELSHIN, S. (2005a) Syntactic and Semantic Decomposition Strategies for Question Answering from Multiple Resources. *The AAAI 2005 Workshop on Inference for Textual Question Answering* Pittsburgh, PA.

KATZ, B., FELSHIN, S., MARTON, G., MORA, F., SHEN, Y. K., ZACCAK, G., AMMAR, A., EISNER, E., TURGUT, A. & WESTRICK, L. B. (2007) CSAIL at TREC 2007 Question Answering. *The Sixteenth Text REtrieval Conference.* NIST.

KATZ, B., LIN, J., STAUFFER, C. & GRIMSON, E. (2003) Answering Questions about Moving Objects in Surveillance Videos. *AAAI Spring Symposium on New Directions in Question Answering.* Palo Alto California.

KATZ, B., MARTON, G., BORCHARDT, G., BROWNELL, A., FELSHIN, S., LORETO, D., LOUIS-ROSENBERG, J., LU, B., MORA, F., STILLER, S., UZUNER, O. & WILCOX, A. (2005b) External Knowledge Sources for Question Answering. *The 14th Annual Text REtrieval Conference (TREC 2005).*

KAY , J. & HANLEY, J. R. (1999) Person-Specific Knowledge And Knowledge Of Biological Categories *Cognitive Neuropsychology,* 16, 171-180.

KEGEL Kegel - Oldest Harley Dealer, http://kegelmotorcycles.com/.

KEIZER, S., OP DEN AKKER, R. & NIJHOLT, A. (2002) Dialogue Act Recognition with Bayesian Networks for Dutch Dialogues. *Third SIGdial Workshop on Discourse and Dialogue.* Philadelphia

KEMMERER, D. (2006) The semantics of space: Integrating linguistic typology and cognitive neuroscience. *Neuropsychologia* 44**,** 1607–1621.

KENNEDY, A. & SZPAKOWITZ, S. (2008) Evaluating Roget's Thesauri. *ACL-08 HLT.* Columbus, Ohio.

KHADER, P., SCHERAG, A., STREB, J. & ROLER, F. (2003) Differences between noun and verb processing in a minimal phrase context: a semantic priming study using event-related brain potentials. *Cognitive Brain Research,* 17**,** 293-313.

KILGARRIFF, A. (1998) Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs, *Computer Speech and Language, Special Issue on Evaluation.,* 12.

KIMURA, Y., ARAKI, K. & TOCHINAI, K. (2007) Identification of Spoken Questions Using Similarity-Based TF.AoI. *Systems and Computers in Japan,* 38**,** 81-94.

KLEIN, D. & MURPHY, G. (2001) The Representation of Polysemous Words. *Journal of Memory and Language,* 45**,** 259–282.

KLEIN, D. & MURPHY, G. (2002) Paper has been my ruin: conceptual relations of polysemous senses. *Journal of Memory and Language,* 47**,** 548-570.

KOPP, S., GESELLENSETTER, L., KRÄMER, N. C. & WACHSMUTH, I. (2005) A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application, *Lecture Notes in Computer Science*

KRAL, P., CERISARA, C. & KLECKOVA, J. (2006) Automatic Dialog Acts Recognition Based On Sentence Structure. *International Conference on Acoustics, Speech, and Signal Processing ICASSP'06.* Toulouse, France.

KRAL, P., CERISARA, C. & KLECKOVA, J. (2007) Lexical Structure for Dialogue Act Recognition. *JOURNAL OF MULTIMEDIA,* 2**,** 1-8.

KRÄMER, N. C., SIMONS, N. & KOPP, S. (2007) The Effects of an Embodied Conversational Agent's Nonverbal Behavior on User's Evaluation and Behavioral Mimicry *Lecture Notes In Artificial Intelligence,* 4722**,** 238 - 251

KUPERBERG, G. R., LAKSHMANAN, B. M., CAPLAN, D. N. & HOLCOMB, P. J. (2006) Making sense of discourse: An fMRI study of causal inferencing across sentences. *NeuroImage* 33 343–361.

LAHAM, D. (1998) Latent Semantic Analysis @ CU Boulder. Boulder Colorado.

LAMEL, L., BENNACEF, S., GAUVAIN, J. L., H., D. & N., T. J. (2002) User evaluation of the MASK kiosk. *Speech Communication,* 38**,** 131 - 139.

LAMEL, L., BENNACEF, S., ROSSET, S., DEVILLERS, L., FOUKIA, S., GANGOLF, J. J. & J_L., G. (1997) The LIMSI RailTel System: Field trial of a telephone service for rail travel information. *Speech Communication,* 23**,** 67-82

LANDAUER, T. K., FOLTZ, P. W. & LAHAM, D. (1998) An Introduction to Latent Semantic Analysis. *Discourse Processes,* 25 259-284.

LANGER, H., LUNGEN, H. & BAYERL, P. S. (2004) Text Type Structure and Logical Document Structure. *the ACL 2004 Workshop on Discourse Annotation.* Barcelona, Spain.

LAWLESS, H. & CAIN, W. (1975) Recognition memory for odors *Chem. Sens. Flavour,* 1**,** 331-337.

LAWS, K. R., GALE, T. M., LEESON, V. C. & CRAWFORD, J. R. (2004) WHEN IS CATEGORY SPECIFIC IN ALZHEIMER'S DISEASE? *Cortex* 41**,** 001-012.

LE BIGOT, L., JAMET, E. & ROUET, J.-F. (2004) Searching information with a natural language dialogue system: a comparison of spoken vs. written modalities. *Applied Ergonomics,* 35**,** 557–564.

LEE, S. (1994) *Untersuchungen zur Valenz des Adjektivs in der deutschen Gegenwartssprache*, Berlin: Lang

LEE, Y., LEE, H. & GORDON, P. C. (2007) Linguistic complexity and information structure in Korean: Evidence from eye-tracking during reading. *Cognition* 104 495–534.

LEFEVRE, F., GASIC, M., JURCICEK, S., KEIZER, S., MAIRESSE, F., THOMSON, B., YU, K. & YOUNG, S. (2009) k-Nearest Neighbor Monte-Carlo Control Algorithm for POMDP-Based Dialogue Systems. IN HEALEY, P., PIERACCINI, R., BYRON, D., YOUNG, S. & PURVER, M. (Eds.) *The SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* London, UK.

LESCH, S., KLEINBAUER, T. & ALEXANDERSSON, J. (2005) A new Metric for the Evaluation of Dialog Act Classication. *Dialor05, the Ninth Workshop On The Semantics And Pragmatics Of Dialogue (SEMDIAL 2005).* Nancy, France.

LEVENE, H. (1960) Robust tests for equality of variances IN OLKIN, I. (Ed.) *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling.* Stanford University Press.

LEVIN, B. (1993) *English verb classes and alternations: a preliminary investigation, ,* Chicago, University of Chicago Press.

LEVIN, L., LANGLEY, C., LAVIE, A., GATES, D., WALLACE, D. & PETERSON, K. (2003) Domain Specific Speech Acts for Spoken Language Translation. *4th SIGdial Workshop on Discourse and Dialogue.* Sapporo, Japan.

LI, W. (1992) Random Texts Exhibit Zipf's-Law-Like Word Frequency Distributions. *IEEE Transactions on Information Theory,* 38**,** 1842-1845.

LI, Y., BANDAR, Z. & MCLEAN, D. (2003) An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering,* 15.

LI, Y., BANDAR, Z., MCLEAN, D. & O'SHEA, J. (2004) A method for measuring sentence similarity and its application to conversational agents. IN BARR, V. & MARKOV, Z. (Eds.) *The 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004),*. Miami Beach, FL. , AAAI Press.

LI, Y., BANDAR, Z., MCLEAN, D. & O'SHEA, J. (2006) Sentence Similarity Based on Semantic Nets and Corpus Statistics. *IEEE Transactions on Knowledge and Data Engineering,* 18**,** 1138-1150.

LIANG, J., LI, C. & MEHROTRA, S. (2003) Efficient Record Linkage in Large Data Sets *the 8th International Conference on Database Systems for Advanced Applications (DASFAA 2003)* Kyoto, Japan.

LIN, D. (1998) An Information-Theoretic Definition of Similarity. *International Conference on Machine Learning.* Madison, Wisconsin.

LISA (2009) English Abstract Nouns - English Grammar. English the Easy Way., http://english-the-easy-way.com/

LITMAN, D. J. (1996) Cue Phrase Classification Using Machine Learning. *J. Artif. Intell. Res. (JAIR),* 5**,** 53-94.

LITMAN, D. J. & PAN, S. (2002) Designing and Evaluating an Adaptive Spoken Dialogue System

User Modeling and User-Adapted Interaction, 12**,** 111-137.

LITTLE, W., FOWLER, H. W. & COULSON, J. (1983) *The Shorter Oxford English Dictionary,* London, Book Club Associates.

LIU, X., ZHOU, Y. & ZHENG, R. (2007a) Sentence Similarity based on Dynamic Time Warping, *International Conference on Semantic Computing, ICSC 2007*

LIU, Y. (2006) Using SVM and Error-correcting Codes for Multiclass Dialog Act Classification in Meeting Corpus. *Proceedings of Interspeech - ICSLP.*

LIU, Y., ZHANG, D., LU, G. & MA, W.-Y. (2007b) A survey of content-based image retrieval with high-level semantics. *Pattern Recognition* 40 262 – 282.

LOCKE, W. N. & BOOTH, A. D. (Eds.) (1955) *Machine Translation of Languages.*, John Wiley & Sons, New York.

LODHI, H., SAUNDERS, C., SHAWE-TAYLOR, J., CRISTIANINI, N. & WATKINS, C. (2002) Text Classification using String Kernels. *The Journal of Machine Learning Research* 2**,** 419 - 444.

LONGACRE, R. (1976) *An anatomy of speech notions*, Peter de Ridder Press.

LOUWERSE, M. & CROSSLEY, S. (2006) Dialog Act Classification Using N-Gram Algorithms. *The Nineteenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2006.*

LU, L. H., CROSSON, B., NADEAU, S. E., HEILMAN, K. M., GONZALEZ-ROTHI, L. J., RAYMERD, A., GILMORE, R. L., BAUER, R. M. & ROPER, S. N. (2002) Category-specific naming deficits for objects and actions: semantic attribute and grammatical role hypotheses. *Neuropsychologia,* 40**,** 1608–1621.

MAHON, B. Z. & CARAMAZZA, A. (2003) There are facts…and then there are facts. *TRENDS in Cognitive Sciences,* 7.

MALETIC, J. I. & MARCUS, A. (2000) Data Cleansing: Beyond Integrity Analysis. *IQ2000.*

MANNING, C. D., RAGHAVAN, P. & SCHÜTZE, H. (Eds.) (2008) *Introduction to Information Retrieval*, Cambridge University Press.

MARSH, J. E., HUGHES, R. W. & JONES, D. M. (2008) Auditory distraction in semantic memory: A process-based approach. *Journal of Memory and Language,* 58**,** 682–700.

MARTIN, R. C. (2003) Language Processing: Functional Organization and Neuroanatomical Basis. *Annual Review of Psychology,* 54**,** 55-89.

MARTON, G. (2006) Nuggeteer: Automatic Nugget-Based Evaluation using Descriptions and Judgements. *Human Language Technology Conference HLT-NAACL.* New York University.

MASSARO, D. W., COHEN, M. M., BESKOW, J. & COLE, R. A. (2000) Developing and evaluating conversational agents. IN CASSELL, J., SULLIVAN, J., PREVOST, S. & CHURCHILL, E. (Eds.) *Embodied conversational agents.* Cambridge, MA, MIT Press.

MASTERMAN, M., NEEDHAM, R. M. & SPÄRCK JONES, K. (1959) The Analogy Between Mechanical Translation and Library Retrieval, *International Conference on Scientific Information National Academy of Sciences—National Research Council.*

MATTHEWS, A., ANDERSON, N., ANDERSON, J. & JACK, M. A. (2008) The effects of personality and individualised product portrayals in the usability of 3D embodied conversational agents in an eBanking scenario *International Conference on Intelligent Virtual Agents, (IVA-08).* Tokyo, Japan.

MCCRAE, R. R. & COSTA, P. T. (1987) Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology,* 52**,** 81-90.

MCGEARY, Z., BAYRIAMOVA, Z., MATIESANU, C. & DANIELS, D. (2005) Online Self-service: The Slow Road to Search Effectiveness. *Customer Relationship Management.*

MCGOLDRICK, P. J., KEELING, K. A. & BEATTY, S. F. (2008) A typology of roles for avatars in online retailing *Journal of Marketing Management, ,* Volume 24**,** 433-461.

MCKEVITT, P., D., P. & WILKS, Y. (1999) Why machines should analyse intention in natural language dialogue. *Int. J. Human-Computer Studies,* 51**,** 947-989.

MCRAE, K. & CREE, G. S. (2002) Factors underlying category-specific semantic deficits. IN FORDE, E. M. E. & HUMPHREYS, G. W. (Eds.) *Category specificity in brain and mind.* Psychology Press.

MEGURO, T., HIGASHINAKA, R., DOHSAKA, K., MINAMI, Y. & H., I. (2009) Analysis of Listening-oriented Dialogue for Building Listening Agents. IN HEALEY, P., PIERACCINI, R., BYRON, D., YOUNG, S. & PURVER, M. (Eds.) *The SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* London, UK.

MICHIE, D. (1993) Turing's Test and conscious thought. *Artificial Intelligence,* 60**,** 1-22.

MICHIE, D. (2001) Return of the Imitation Game. *Electronic Transactions in Artificial Intelligence.*

MIHALCEA, R., CORLEY, C. & STRAPPARAVA, C. (2006) Corpus-based and Knowledge-based Measures of Text Semantic Similarity. *Proceedings of the American Association for Artificial Intelligence (AAAI 2006).* Boston.

MIHALCEA, R. & PEDERSEN, T. (2005) Advances in Word Sense Disambiguation. *AAAI-2005.*

MILLER, G. A. (1972) English verbs of motion: a case study in semantics and lexical memory. IN MELTON, A. W. & MARTIN, E. (Eds.) *Coding Processes in Human Memory,* Washington, DC, Wiley.

MILLER, G. A. (1995) WordNet: A Lexical Database for English. *Communications of the ACM* 38**,** 39-4.

MILLER, G. A., BECKWITH, R., FELLBAUM, C., GROSS, D. & MILLER, K. (1990) Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography* 3**,** 235-244.

MILLER, G. A. & CHARLES, W. G. (1991) Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes* 6 1-28.

MIN, F., WENYIN, L. & CHEN, W. (2006) Answer Clustering and Fusion in a User-Interactive QA System. *Second International Conference on Semantics, Knowledge and Grid.*

MINKER, W., BENNACEF, S. & GAUVAIN, J.-L. (1996) A stochastic case frame approach for natural language understanding *Fourth International Conference on Spoken Language, ICSLP 96.* . Philadelphia, PA

MISU, T. & KAWAHARA, T. (2010) Bayes risk-based dialogue management for document retrieval system with speech interface. *Speech Communication,* 52**,** 61-71.

MONTEMAGNI, S. & PIRRELLI, V. (1998) Augmenting WordNet-like lexical resources with distributional evidence. An application-oriented perspective. *Use of WordNet in Natural Language Processing Systems.*

MORI, J., PRENDINGER, H. & ISHIZUKA, M. (2003) Evaluation of an embodied conversational agent with affective behavior. . *The 2nd International Joint Conference on Autonomous Agents and Multiagent Systems.* Melbourne, Australia.

MOSS, H. E. & TYLER, L. K. (2000) A progressive category-specific semantic deficit for non-living things, *Neuropsychologia* 38**,** 60-82.

NAUMANN, K. (2000) Classification of Adjectives.

NEWTON, P. K. & BARRY, C. (1997) Concreteness Effects in Word Production but Not Word Comprehension in Deep Dyslexia *Cognitive Neuropsychology,* 14**,** 481 - 509

NGAI, G. & YAROWSKY, D. (2000) Rule Writing or Annotation: Cost-efficient Resource Usage for Base Noun Phrase Chunking. *38th Annual Meeting of the Association for Computational Linguistics* Hong Kong, China.

NIETZSCHE, F. (1888) *Twilight of the Idols.*

NIST PRISE Search Design Notes. NIST.

NORMAN, W. (1963) Toward an adequate taxonomy of personal personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of abnormal and social psychology,* 66**,** 574-583.

O'DONAILL, E. & NI CHURRAIGHIN, D. (1995) *Now You're Talking: Multi-Media Course in Irish for Beginners*, Gill & Macmillan Ltd.

O'SHEA, J. D. BANDAR, Z., CROCKETT, K. & MCLEAN, D. (2010) Benchmarking Short Text Semantic Similarity. *Int. J. Intelligent Information and Database Systems,* 4**,** 103-120.

O'SHEA, J. D., BANDAR, Z., CROCKETT, K. & MCLEAN, D. (2008) A Comparative Study of Two Short Text Semantic Similarity Measures. *Lecture Notes in Artificial Intelligence*.

O'SHEA, K., BANDAR, Z. & CROCKETT, K. (2009) Towards a New Generation of Conversational Agents Based on Sentence Similarity *Lecture Notes Electrical Engineering,* 39**,** 505-514.

O'SHEA, K., BANDAR, Z. & CROCKETT, K. (2008) A Novel Approach for Constructing Conversational Agents using Sentence Similarity Measures. *Proceedings of the World Congress on Engineering WCE 2008.* London, U.K.

OGDEN, C. (1930) *Basic English: A General Introduction with Rules and Grammar.,* London, Paul Treber & Co., Ltd. .

OLIVEIRA, E. & SARMENTO, L. (2002) Emotional Valence-Based Mechanisms and Agent Personality. *Lecture Notes In Computer Science,* 2507**,** 152 - 162

OP DEN AKKER, H. & OP DEN AKKER, R. (2009) Are You Being Addressed? - real-time addressee detection to support remote participants in hybrid meetings. IN HEALEY, P., PIERACCINI, R., BYRON, D., YOUNG, S. & PURVER, M. (Eds.) *The SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* London, UK.

OPPENHEIM, A. N. (1992) *Questionnaire Design, Interviewing and Attitude Measurement*, Continuum

ORTONY, A., CLORE, G. L. & COLLINS, A. (1990) *The Cognitive Structure of Emotions,* Cambridge University Press.

OWDA, M., BANDAR, Z. & CROCKETT, K. (2007) Conversation-Based Natural Language Interface to Relational Databases. *IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops.*

PALMER, F. (1988) *The English Verb*, Longman Linguistics Library.

PALMER, M. (2006) Introduction to Computational Linguistics. University of Colorado.

PARK, S., KIM, H., KO, Y. & SEO, J. (2000) Implementation of an efficient requirements-analysis supporting system using similarity measure techniques. *Information and Software Technology,* 42**,** 429–438.

PEARSALL, J. (2001) *Oxford Concise English Dictionary*, Oxford University Press.

PEXMAN, P. M., HARGREAVES, I. S., EDWARDS, J. D., HENRY, L. C. & GOODYEAR, B. G. (2007) Neural Correlates of Concreteness in Semantic Categorization. *Journal of Cognitive Neuroscience* 19**,** 1407-1419

PHILLIPS, J. A., NOPPENEY, U., W., H. G. & PRICE, C. J. (2002) Can segregation within the semantic system account for category-specific deficits? . *Brain,* 125**,** 2067-2080.

PILATO, G., AUGELLO, A., VASSALLO, G. & GAGLIO, S. (2009) Geometric Algebra Rotors for Sub-symbolic Coding of Natural Language Sentences *Lecture Notes in Computer Science,* 4692/2009**,** 42-51.

PLANTEC, P. (1998) The Zen of Scripting Verbots. Virtual Personalities Inc.

POURATIAN, N., BOOKHEIMER, S. Y., RUBINO, R., MARTIN, N. A. & TOGA, A. W. (2003) Category-specific naming deficit identified by intraoperative stimulation mapping and postoperative neuropsychological testing. *Journal of neurosurgery,* 99**,** 170-176.

PRASAD, R. & WALKER, W. (2002) Training a Dialogue Act Tagger For Human-Human and Human-Computer Travel Dialogues. *The 3rd SIGdial workshop on Discourse and dialogue.* Philadelphia, Pennsylvania.

PRINCETON WordNet Search - 3.0. . Princeton University Computer Science Laboratory.

PU, P. & CHEN, L. (2006) Trust building with explanation interfaces. *The 11th international conference on Intelligent user interfaces.* Sydney, Australia

PUPPALA, P., THAKORE, H. & EDELMAN, M. J. (2009) Case Report of Mesial Temporal Sclerosis With Seizures and Psychosis: An Interface Between Psychiatry and Neurology. *Prim Care Companion J Clin Psychiatry,* 11**,** 37–38.

QUARTERONI, S. & MANANDHAR, S. (2006) Incorporating User Models in Question Answering to Improve Readability. *KRAQ'06* Trento, Italy

QUINLAN, J. R. (1986) Induction of Decision Trees. *Machine Learning,* 1**,** 81-106.

QUINLAN, J. R. (1993) *C4.5: programs for machine learning.,* San Mateo, California, Morgan Kaufmann Publishers.

QUIRK, R., GREENBAUM, S., LEECH, G. & SVARTIK, J. (1985) *A Comprehensive Grammar of the English Language,* Harlow, Addison Wesley Longman Ltd.

RACHIDI, R. (1989) *Contrary relations in the field of German adjectives*, Max Niemeyer Verlag

RACTER (1984) *The Policeman's beard is Half Constructed*, Warner Books.
RALLABANDI, V. P. S. & SETT, S. K. (2008) Knowledge-based image retrieval system. *Knowledge-Based Systems* 21**,** 89–100.
RAVI, S., KNIGHT, K. & SORICUT, R. (2008) Automatic prediction of parser accuracy. *The Conference on Empirical Methods in Natural Language Processing* Honolulu, Hawaii, Association for Computational Linguistics Morristown, NJ, USA.
RESNIK, P. (1999) Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research,* Vol. 11**,** 95-130.
RESNIK, P. & DIAB, M. (2000) Measuring Verb Similarity. *Twenty Second Annual Meeting of the Cognitive Science Society (COGSCI2000).* Philadelphia.
RICE, G. (2006) Are Viruses Alive? , Montana State University.
RICE, J. A. (1994) *Mathematical Statistics and Data Analysis*, Duxbury Press.
RIECK, K. & LASKOV, P. (2007) Linear-Time Computation of Similarity Measures for Sequential Data. *Advances in Neural Information Processing Systems* 19, **,** 1177–1184.
ROBINSON, S., TRAUM, D., ITTYCHERIAH, M. & HENDERER, J. (2006) "What Would You Like to Talk About?" An Evaluation of Social Conversations with a Virtual Receptionist. *Lecture Notes in Computer Science,* 4133/2006**,** 1125-1131.
RODRIGUEZ, A. & EGENHOFER, M. J. (2003) Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering,* 5 442 - 456.
ROGERS, C. R. (1951) *Client-Centered Therapy: Its Current Practice, Implications, and Theory* Trans-Atlantic Publications (December 1995).
ROGERS, T. T. & PLAUT, D. C. (2002) Connectionist perspectives on category specific deficits. IN FORDE, E. M. E. & HUMPHREYS, G. W. (Eds.) *Category specificity in brain and mind.* Psychology Press.
ROSSELL, S. L., SHAPLESKE, J. & DAVID, A. S. (1988) Sentence verification and delusions: a context specific deficit. *Psychological medicine,* 28**,** 1189-1198.
ROTHWELL, J., BANDAR, Z., O'SHEA, J. & MCLEAN, D. (2006) Silent Talker: A New Computer-Based System for the Analysis of Facial Cues to Deception. *Applied Cognitive Psychology,* 20**,** 757–777.
ROUSSINOV, D. & ZHAO, J. L. (2003) Automatic discovery of similarity relationships through Web mining. *Decision Support Systems* 35**,** 149-166.
RUBENSTEIN, H. & GOODENOUGH, J. (1965) Contextual Correlates of Synonymy. *Communications of the ACM,* 8**,** 627-633.
SAGAE, K., DAVIS, E., LAVIE, A., MACWHINNEY, B. & WINTNER, S. (2007) High-accuracy annotation and parsing of CHILDES transcripts. *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition* Prague, Czech Republic Association for Computational Linguistics Morristown, NJ, USA
SAGAE, K., GWEN CHRISTIAN, G., DEVAULT, D. & DAVID R. TRAUM, D. R. (2009) Towards Natural Language Understanding of Partial Speech Recognition Results in Dialogue Systems. *NAACL HLT 2009.* Boulder, Colorado,, Association for Computational Linguistics.
SAGGION, H., BONTCHEVA, K. & CUNNINGHAM, H. (2003) Robust Generic and Query-based Summarisation. *The tenth conference on European chapter of the Association for Computational Linguistics.* Budapest, Hungary.
SAHAMI, M. (2006) Mining theWeb to Determine Similarity Between Words, Objects, and Communities. *Flairs*
SALEEM, K., BELLAHSENE, Z. & HUNT, E. (2008) PORSCHE: Performance ORiented SCHEma mediation. *Information Systems,* 33**,** 637-657.

SALOMON, J., KING, S. & OSBORNE, M. (2002) Framewise Phone Classification Using Support Vector Machines. *7th International Conference on Spoken Language Processing (ICSLP2002 - INTERSPEECH 2002)* Denver, Colorado, USA.

SALTON, G. & ALLAN, J. (1993) Selective Text Utilization and Text Traversal. *Hypertext '93.*

SALTON, G., WONG, A. & YANG , C. S. (1975) A Vector Space Model for Automatic Indexing. *Communications of the ACM,* 18**,** 613–620.

SAMMUT, C. (2001) Managing Context in a Conversational Agent. *Electronic Transactions in Artificial Intelligence.*

SANDERS, G. A. & SCHOLTZ, J. (2000) Measurement and Evaluation of Embodied Conversational Agents. . IN CASSELL, J., SULLIVAN, J., PREVOST, S. & CHURCHILL, E. (Eds.) *Embodied Conversational Agents.* MIT Press.

SANTOS, L. R. & CARAMAZZA, A. (2002) The domain-specific hypothesis. IN FORDE, E. M. E. & HUMPREYS, G. W. (Eds.) *Category specificity in brain and mind.* Psychology Press.

SARTORI, G., MAMELI, F., POLEZZI, D. & LOMBARDI, L. (2006) An ERP study of low and high relevance semantic features *Brain Research Bulletin,* 69**,** 182-186

SARTORI, G., MIOZZO, M. & JOB, R. (1993) Category-specific naming impairments? Yes. . *Quarterly Journal of Experimental Psychology,* 46A.

SAUCIER, G. & GOLDBERG, L. R. (1996) The language of personality: Lexical perspectives on the five-factor model. IN WIGGINS, J. S. (Ed.) *The five-factor model of personality: Theoretical perspectives* New York, Guilford.

SCHMITT, A., HEINROTH, T. & LISCOMBE, J. (2009) On NoMatchs, NoInputs and BargeIns: Do Non-Acoustic Features Support Anger Detection? IN HEALEY, P., PIERACCINI, R., BYRON, D., YOUNG, S. & PURVER, M. (Eds.) *The SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* London, UK.

SEARLE, J. R. (1969) *Speech Acts: An Essay in the Philosophy of Language* Cambridge University Press.

SEARLE, J. R. (1980) Minds, brains and programs. *Behavioural and Brain Sciences,* 3**,** 417-424.

SEARLE, J. R. (1999) *Mind, Language and Society*, Weidenfield & Nicholson

SEMENZA, C., MONDINI, S. & CAPPELLETTI, M. (1997) The grammatical properties of mass nouns: An aphasia case study *Neuropsychologia,* 35, 669-675.

SEMERARO, G., ANDERSEN, V., ANDERSEN, H. H. K., DEGEMMIS, M. & LOPS, P. (2008) User profiling and virtual agents: a case study on e-commerce services. *Universal Access in the Information Society (UAIS)* 7**,** 179-194.

SEMERARO, G., ANDERSON, H., ANDERSON, V., LOPS, P. & ABBATTISTA, F. (2003) Evaluation and Validation of a Conversational Agent Embodied in a Bookstore. *Lecture Notes in Computer Science,* 2615**,** 360-371.

SERAFIN, R., DI EUGENIO, B. & GLASS, M. (2003) Latent Semantic Analysis for dialogue act classification. *The 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology.* Edmonton, Canada.

SERENDIP (2003) Parietal Lobe Definiton, http://serendip.brynmawr.edu/bb/kinser/definitions/def-parlobe.html.

SHAW, R., SOLWAY, B., GAIZAUSKAS, R. & GREENWOOD, M. A. (2008) Evaluation of Automatically Reformulated Questions in Question Series. *Coling 2008* Manchester.

SHOKOUFANDEH, A., BRETZNER, L., MACRINI, D., DEMIRCI, M. F., JONSSON, C. & DICKINSON, S. (2006) The representation and matching of categorical shape. *Computer Vision and Image Understanding* 103 139–154.

SIGDIAL (2009) Proceedings of the SIGDIAL 2009 Conference. IN HEALEY, P., PIERACCINI, R., BYRON, D., YOUNG, S. & PURVER, M. (Eds.) *The SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* London, UK.

SILVERI, M. C., PERRI, R. & CAPPA, A. (2003) Grammatical class effects in brain-damaged patients: Functional locus of noun and verb deficit. *Brain and Language* 85**,** 49–66.

SINCLAIR, J. (2001) *Collins Cobuild English Dictionary for Advanced Learners*, HarperCollins.

SKANTZE, G. & GUSTAFSON, J. (2009) Attention and Interaction Control in a Human-Human-Computer Dialogue Setting. IN HEALEY, P., PIERACCINI, R., BYRON, D., YOUNG, S. & PURVER, M. (Eds.) *The SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* London, UK.

SLEVC, L. R. & FERREIRA, V. S. (2006) Halting in single word production: A test of the perceptual loop theory of speech monitoring. *Journal of Memory and Language,* 54**,** 515–540.

SMUCKER, M. D., ALLAN, J. & BLAGOVEST DACHEV, B. (2007) UMass Complex Interactive Question Answering (ciQA) 2007: Human Performance as Question Answerers. *The Sixteenth Text REtrieval Conference.* NIST.

SNEDECOR, G. W. & COCHRAN, W. G. (1989) *Statistical Methods*, Iowa State University Press.

SNOW, C. P. (1963) *The Two Cultures* Cambridge University Press.

SPÄRCK-JONES, K. (1972) A Statistical Interpretation of Term Specificity and its Application in Retrieval *Journal of Documentation,* vol. 28**,** 11–21.

STEWART, F., PARKIN, A. J. & HUNKIN, N. M. (1992) Naming impairments following recovery from herpes simplex encephalitis: Category-specific?, Quarterly Journal of Experimental Psychology**,** 261-284.

STOILOS, G., STAMOU, G. & KOLLIAS, S. (2005) A String Metric for Ontology Alignment. *4th International Semantic Web Conference (ISWC 2005).* Galway.

STOLCKE, A., RIES, K., COCCARO, N., SHRIBERG, E., BATES, R., JURAFSKY, D., TAYLOR, P., MARTIN, R., VAN ESS-DYKEMA, C. & METEER, M. (2000) Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics* 26**,** 339-373.

SU, X. & GULLA, J. A. (2006) An information retrieval approach to ontology mapping. *Data & Knowledge Engineering,* 58**,** 47–69.

THIERRY, G., D., C. & DEMONET, J. F. (2003) Electrophysiological comparison of grammatical processing and semantic processing of single spoken nouns. *Cognitive Brain Research* 17 535–547.

THOMSON, A. J. & MARTINET, A. V. (1969) *A Practical English Grammar*, Oxford University Press.

THOMSON, B., GASIC, M., KEIZER, S., MAIRESSE, F., SCHATZMANN, J., YU, K. & YOUNG, S. (2008) User study of the Bayesian Update of Dialogue State approach to dialogue management. *Interspeech 2008.* Brisbane, Australia.

THOMSON, W. (1889) Electrical Units of Measurement. *Popular Lectures and Addresses,* 1.

THURSTONE, L. L. (1927) Psychophysical analysis. *American Journal of Psychology,* 38**,** 68-89.

TRAUM, D. & HABASH, N. (2000) Generation from Lexical Conceptual Structures. *Workshop on Applied Interlinguas, ANLP-2000.* . Seattle, WA

TRAUM, D. R. & HINKELMAN, E. A. (1992) Conversation Acts in Task-Oriented Spoken Dialogue *Computational Intelligence,* , 8**,** 575-599.

TSATSARONIS, G., VARLAMIS, I. & VAZIRGIANNIS, M. (2010) Text Relatedness Based on a Word Thesaurus *Journal of Artificial Intelligence Research,* 37.

TUFIS, D., ION, R. & IDE, N. (2004) Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. *The 20th international conference on Computational Linguistics.* Geneva, Switzerland

TURING, A. M. (1948) Intelligent Machinery. National Physical Laboratory.

TURING, A. M. (1950) Computing Machinery and Intelligence. *Mind, New Series,* 59**,** 433-460.

TURING, A. M. (1952) Can Automatic Calculating Machines Be Said To Think? , BBC 3rd programme.

TURNEY, P. D. (2001a) Answering subcognitive Turing Test questions: a reply to French. *J. Expt. Theor. Artificial Intelligence,* 13**,** 409-419.

TURNEY, P. D. (2001b) Mining the Web for Synonyms:PMI-IR versus LSA on TOEFL. *12th European Conference on Machine Learning (ECML-2001).* Freiburg, Germany.

TVERSKY, A. (1977) Features of Similarity. *Psychological Review,* 84 327-352.

UZUNER, O., DAVIS, R. & KATZ, B. (2004) USING EMPIRICAL METHODS FOR EVALUATING EXPRESSION AND CONTENT SIMILARITY *37th International Conference on System Sciences.* Hawaii, IEEE.

VALCOURT, G. & WELLS, L. (1999) *Mastery: A University Word List Reader,* Michigan, The University of Michigan Press

VAN NOORD, G. (1997) An Efficient Implementation of the Head-Corner Parser. *Computational Linguistics,* 23.

VAN OVERSCHELDE, J. P., RAWSON, K. A. & DUNLOSKYC, J. (2004) Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language,* 50**,** 289–335.

VAN VALIN, R. D. (2008) An Overview of Role and Reference Grammar. Buffalo University.

VENKATARAMAN, A., FERRER, L., STOLCKE, A. & SHRIBERG, E. (2003) TRAINING A PROSODY-BASED DIALOG ACT TAGGER FROM UNLABELED DATA. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003*

VENKATARAMAN, A., STOLCKE, A. & SHRIBERG, E. (2002) Automatic Dialog Act Labeling With Minimal Supervision. *9th Australian International Conference on Speech Science and Technology.*

VERBREE, D., RIENKS, R. & HEYLEN, D. (2006) Dialogue-Act Tagging Using Smart Feature Selection; Results On Multiple Corpora. *IEEE Spoken Language Technology Workshop.*

VIGGLIOCCO, G., VINSON, D. P. & SIRI, S. (2005) Semantic similarity and grammatical class in naming actions. *Cognition,* 94**,** B91–B100.

VIGLIOCCO, G., VINSON, D., LEWIS, W. & GARRETT, M. (2002a) Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology,* 48**,** 422-488

VIGLIOCCO, G., VINSON, D. P., ARCIULI, J. & BARBER, H. (2008) The role of grammatical class on word recognition. *Brain and Language* 105 175–184.

VIGLIOCCO, G., VINSON, D. P., DAMIAN, M. F. & LEVELT, W. (2002b) Semantic distance effects on object and action naming. *Cognition,* 85 B61-B69.

VIGLIOCCO, G., VINSON, D. P., LEWIS, W. & GARRETT, M. F. (2004) Representing the meanings of object and action words: The featural and unitary semantic space hypothesis *Cognitive Psychology,* 48**,** 422-488.

VINSON, D. P., VIGLIOCCO, G., CAPPA, S. & SIRI, S. (2003) The breakdown of semantic knowledge: Insights from a statistical model of meaning representation *Brain and Language,* 86**,** 347-365

WALKER, M., ABERDEEN, J., BOLAND, J., BRATT, E., GAROFOLO, J., HIRSCHMAN, L., LE, A., LEE, S., NARAYANAN, S., PAPINENI, K., PELLOM, B., POLIFRONI, J., POTAMIANOS, A., PRABHU, P., RUDNICKY, A., SANDERS, G., SENEFF, S., STALLARD, S. & WHITTAKER, S. (2001) Darpa communicator dialog travel planning systems: The June 2000 data collection. In Submitted to EUROSPEECH 2001. . *EUROSPEECH 2001 7th European Conference on Speech Communication and Technology 2nd INTERSPEECH Event,* Aalborg, Denmark

WALKER, M., HIRSCHMAN, L. & ABERDEEN, J. (2000) Evaluation for Darpa Communicator Spoken Dialogue Systems. . *Language Resources and Evaluation Conference,* Athens, Greece.

WALLACE, R. (2003) The Elements of AIML Style. Alice AI foundation Inc

WALLACE, R. (2009) Alicebot, http://alicebot.blogspot.com/.

WANG, F., NGO, C.-W. & PONGA, T.-C. (2008) Structuring low-quality videotaped lectures for cross-reference browsing by videotext analysis. *Pattern Recognition.*

WARRINGTON, E. K. & MCCARTHY, R. (1983) Category specific access dysphasia. *Brain,* 106 859-878.

WARRINGTON, E. K. & SHALLICE, T. (1984) Category-specific semantic impairments. *Brain,* 107**,** 829-853.

WEBB, N., HEPPLE, M. & WILKS, Y. (2005a) Dialogue Act Classification Based on Intra-Utterance Features. *AAAI 2005.* Pittsburgh, Pennsylvania, AAAI Press.

WEBB, N., HEPPLE, M. & WILKS, Y. (2005b) Error analysis of dialogue act classification. *Proceedings of the 8th International Conference on Text, Speech and Dialogue.* Karlovy Vary, Czech Republic.

WEBB, N. & LIU, T. (2008) Investigating the Portability of Corpus-Derived Cue Phrases for Dialogue Act Classification. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008),* Manchester.

WEBMASTER, B. [bnc] British National Corpus. Oxford University Computing Services.

WEIZENBAUM, J. (1966) ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the ACM,* 9**,** 36-41.

WEST, M. P. (1953) *A General Service List of English Words,* London, Longman.

WHATMOUGH, C. & CHERKOW, H. (2002) Category-specific recognition impairments in Alzheimer's disease. IN FORDE, E. M. E. & HUMPHREYS, G. W. (Eds.) *Category specificity in brain and mind.* Psychology press.

WIEBE, J., BRUCE, R. & O'HARA, T. (1999) Development and use of a gold standard data set for subjectivity classifications. *The 37th Annual Meeting of the Association for Computational Linguistics (ACL-99).*

WITTEN, I. H. & EIBE, F. (2005) *Data Mining: Practical Machine Learning Tools and Techniques,* San Francisco, Elsevier.

WU, C.-W., TSAI, R. T.-H., LEE, C.-W. & HSU, W.-L. (2007) Web taxonomy integration with hierarchical shrinkage algorithm and fine-grained relations. *Expert Systems with Applications,* 35**,** 2123-2131.

WU, Z. & PALMER, M. (1994) Verb Semantics and lexical selection *The 32nd Annual Meeting of the Association for Computational Linguistics,*. Las Cruces, New Mexico.

YANG, F., TUR, G. & SHRIBERG, E. (2008) Exploiting Dialogue Act Tagging And Prosodic Information For Action Item Identification *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.* . Location: Las Vegas, NV.

YEH, J.-Y., KE, H.-R. & YANG, W.-P. (2008) iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network. *Expert Systems with Applications,* 35**,** 1451–1462.

YI, H., RAJAN, D. & CHIA, L.-T. (2006) A motion-based scene tree for browsing and retrieval of compressed videos. *Information Systems* 31 638–658.

YOUNG, S., GASIC, M., KEIZER, S., MAIRESSE, F., SCHATZMANN, J., THOMSON, B. & YU, K. (2010) The Hidden Information State model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech and Language, Special Issue on Evaluation.,* 24**,** 150-174.

YOURDICTIONARY (2009) Abstract Nouns. Burlingame, CA LoveToKnow Corp.

YUAN , X. & CHEE, Y. S. (2005) Design and evaluation of Elva: an embodied tour guide in an interactive virtual art gallery. *Computer Animation and Virtual Worlds,* 16**,** 109 - 119.

ZDRAVKOVA, K. (2000) Conceptual Framework for an Intelligent and ChatterBot. *22nd International Conference Information Technology Interfaces ITI 2000.*

ZELIKOVITZ, S. & HIRSH, H. (2000) Improving Short-Text Classification Using Unlabeled Background Knowledge to Assess Document Similarity. *The Seventeenth International Conference on Machine Learning.* Stanford University Morgan Kaufmann Publishers, San Francisco, US.

ZHANG, R. & ZHANG, Z. M. (2006) BALAS: Empirical Bayesian learning in the relevance feedback for image retrieval. *Image and Vision Computing* 24 211–223.

ZUCCO, M. (2003) Anomalies in Cognition: Olfactory Memory. *European Psychologist,* 8**,** 77-86.

ZUKERMAN, I., YE, P., K., G. K. & E., M. (2009) Towards the Interpretation of Utterance Sequences in a Dialogue System. IN HEALEY, P., PIERACCINI, R., BYRON, D., YOUNG, S. & PURVER, M. (Eds.) *The SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* London, UK.

Appendices

Appendix 1.1: Dialogue Samples from NLP-based systems

Sample dialogue from the towninfo domain used for NLP systems (Lefevre et al., 2009, Young et al., 2010)

U: Hi, I am looking for somewhere to eat.
S: You are looking for a restaurant. What type of food do you like?
U: I'd like an Italian somewhere near the museum.
S: Roma is a nice Italian restaurant near the museum.
U: Is it reasonably priced?
S: Yes, Roma is in the moderate price range.
U: What is the phone number?
S: The number of Roma is 385456.
U: Ok, thank you goodbye.

Sample dialogue from UAH from Universidad de Granada (academic ASR information system)

S1: Welcome to the UAH system. How can I help you?
U1: I want to know information about subjects in Computer Science
S2: Which subject do you want to consult?
U2: Language Processors I
S3: What would you like to know?
U3: Can you inform me about the schedules of Group D?
S4: The schedule of Group D is: Mondays from 9 to 10 and Wednesdays from 9 to 11 and from 11 to 13. Anything else?
U4: Can you tell me the name of the lecturer?
S5: The name of the lecturer of the Group D of Language Processors I is David Griol. Anything else?
U5: No, thank you.
S6: Thank you for using the UAH system.

Appendix 2.1 Ethics statement (for Experiments 1, 2.1 and 2.4)

Call for participants in an experimental study of sentence semantic similarity.

We would like to request your participation in a scientific study of semantic similarity. For ethical reasons we are required to ask your permission in advance and let you know what you are agreeing to. We have provided the answers to the key ethical questions below. If you require any further information before agreeing to participate please contact Jim O'Shea (j.d.oshea@mmu.ac.uk, 0161 247 1546)

What will you ask me to do?
If you agree you will be asked to complete a questionnaire in which you will rate the similarity of meaning of 65 pairs of sentences.
You will also be asked to complete a few questions about yourself.
These are your name, age and highest level of qualification. You will also be asked to confirm that you are a native speaker of English (i.e. someone for whom it is their first language, spoken since birth).
We ask for some personal data because sometimes scientific studies produce surprising results which need to be analysed and this background information could help.

Is there any risk?
The sentences do not contain any words which would generally be considered to be offensive. The risk involved is equivalent to looking up an ordinary word in a dictionary.

How long will the data be kept for?
The answers to the questions about yourself will be kept for no longer than is necessary to check for errors or interesting properties of the data. This will be for no longer than 3 months after the first results are published.
The ratings you provide will be separated from the personal data and kept permanently. This is because the data can be very useful in long-term studies. Data of this type collected in the 1960s is still used widely today.

Will you publish my personal information?
We will never disclose your personal information to anyone outside the project. Statistical summaries of the ratings will be published on an international scale. The set of individual ratings, made anonymous by removing personal details, may be made available to bona fide researchers on request.

Appendix 2.2 Instructions (for Experiments 1, 2.1 and 2.4)

Survey: Sentence Semantic Similarity

Thank-you for volunteering to take part in this study.

You may still withdraw before starting the questionnaire or at any point while completing the questionnaire.

Please complete the questionnaire in a single sitting (**you only have to read 2 sentences and write 1 number on each page**).

Please start at the first page and work through in the pages in order from start to end, do not go back and make changes to earlier responses.

Please note that the study does not evaluate you in any way – there are no "right" or "wrong" answers, except in the sense that the right answer to each question is an accurate expression of your personal opinion.

Please read the sentence pairs on the following pages carefully and rate how similar their meanings are.

You should rate the similarity by writing a number between 0.0 (minimum similarity) and 4.0 (maximum similarity) on the form, please do not use values greater than 4.0.
You can use the first decimal place (e.g. 2.2) to show finer degrees of similarity.

You may or may not agree with the statements made in the sentences, but please do not allow this to influence your judgement.

If you have any problems, questions or comments please speak to one of the investigators.

Appendix 2.3 Sample question sheet (for experiments 1 and 2.1)

Please read the pair of sentences below and rate how similar they are in meaning:

Sentence 1	Cord is strong, thick string.
Sentence 2	A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.

The rating scale runs from 0.0 (minimum similarity) to 4.0 (maximum similarity), please do not use values greater than 4.0.

Write your answer in the box on the right	

0.0 The sentences are unrelated in meaning.

1.0 The sentences are vaguely similar in meaning.

2.0 The sentences are very much alike in meaning.

3.0 The sentences are strongly related in meaning.

4.0 The sentences are identical in meaning.

You can use the first decimal place, for example if you think the similarity is half way between 3.0 and 4.0 you can use a value like 3.5.

Appendix 2.4 Personal Data Sheet (for Experiments 1, 2.1, 2.2, 2.3, 2.4)

And now, just a few items of personal information:

Your name (print)	
Your age	
Your highest educational qualification (including subject)	
Confirmation that you are a native English speaker* - please tick	

Signature:

* Native English speaker means that it is your first language and you have spoken it since birth.

Please return to:

Jim O'Shea
Department of Computing and Mathematics
Manchester Metropolitan University
Chester St.
Manchester
M1 5GD

0161 247 1541
j.d.oshea@mmu.ac.uk

Appendix 2.5 Instructions (for Experiments 2.2 and 2.3)
Survey: Sentence Semantic Similarity

Thank-you for volunteering to take part in this study.

You may still withdraw before starting the task or at any point while doing it.

Please complete the task in a single sitting.

You are provided with an envelope containing a set of cards and a recording sheet to write your judgements on (please don't write anything on the cards). The cards have been shuffled into a random order.

Each card has two sentences written on it. Please start by reading through the cards in the order you got them in, thinking about the similarity of the meanings of the two sentences on each card.

Now please sort the cards in a rough order of the similarity of meaning of the sentence pairs. Don't spend too long over this as it is only to help with the final stage.

Finally, please rate the similarity of each pair of sentences by writing a number between 0.0 (minimum similarity) and 4.0 (maximum similarity) on the recording sheets, please do not use values greater than 4.0. You can use the first decimal place (e.g. 2.2) to show finer degrees of similarity. There are some guidance notes on the recording sheets to help you.

You may or may not agree with the statements made in the sentences, but please do not allow this to influence your judgement.

If you have any problems, questions or comments please speak to one of the investigators.

Please note that the study does not evaluate you in any way – there are no "right" or "wrong" answers, except in the sense that the right answer to each question is an accurate expression of your personal opinion.

Appendix 2.6 Sample card (for Experiments 2.2 and 2.3)

SP10

Sentence 1	Cord is strong, thick string.
Sentence 2	A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.

Appendix 2.7 Sample rating sheet (with semantic anchors for Experiment 2.3)

Similarity Rating Sheet

Please enter a rating for the similarity of meaning of each sentence pair.

The rating scale runs from 0.0 (minimum similarity) to 4.0 (maximum similarity), please do not use values greater than 4.0.

SP01	
SP02	
SP03	
SP04	
SP05	
SP06	
SP10	
SP11	
SP12	
SP13	
SP14	
SP15	
SP16	

SP20	
SP21	
SP22	
SP23	
SP24	
SP25	
SP26	
SP30	
SP31	
SP32	
SP33	
SP34	
SP35	
SP36	

SP40	
SP41	
SP42	
SP43	
SP44	
SP45	
SP46	
SP50	
SP51	
SP52	
SP53	
SP54	
SP55	
SP56	

Guidance Notes

If you have difficulty making an assessment here are some descriptions of the main points on the scale to help you:

0.0 The sentences are unrelated in meaning.

1.0 The sentences are vaguely similar in meaning.

2.0 The sentences are very much alike in meaning.

3.0 The sentences are strongly related in meaning.

4.0 The sentences are identical in meaning.

You can use the first decimal place, for example if you think the similarity is half way between 3.0 and 4.0 you can use a value like 3.5.

Appendix 2.8 Sample instruction sheet (without semantic anchors for experiment 2.4)

Please read the pair of sentences below and rate how similar they are in meaning:

Sentence 1	Cord is strong, thick string.
Sentence 2	A smile is the expression that you have on your face when you are pleased or amused, or when you are being friendly.

The rating scale runs from 0.0 (minimum similarity) to 4.0 (maximum similarity), please do not use values greater than 4.0.

Write your answer in the box on the right	

Appendix 3.1 Blocked design for distribution to materials to participants (Experiment 3.1)

Block A1	Block A2	Block B1	Block B2
7 Nouns variety (AQu) conflict (ASt) ear (LBBoP) cat (LBA) weapon (NNMiAr) heap (CNNLF) key (NNTMHe/Ho)	7 Nouns delay (ARLFPoV) pleasure (AEm+) lemon (BFrV) parent (LBR) coffee(NBFo) boat (NNVeOut) diamond (NNGLF)	7 Nouns nation (AIF) doubt (AEm-) rose (BPFlLF) fly (LBI) chair (NNFuIn) hill (NNONMM) fire(NNRPoVb)	7 Nouns temperature (ASc) prejudice (ARLF) hen (LBBiLF) shoe (NNCl) guitar (NNMuILF) crowd (CLB) quay (NNMiArHoHo)
4 Verbs hope 32.2 (s) laugh 40.2 (a) relax 31.1 (s) (lf) climb 51.3.2 (a)(SSC)	3 Verbs change 41.1.1 (p) drink 39.1 (ap)(SC) punish 33 (ap)	4 Verbs cover 47.8 (s) run 51.3.2 (a)(SSC) bargain 36.1 (a) (lf) eat 39.1 (ap)(SC)	3 Verbs dry 45.4 (p) (lf) (PoA) delay 53.1 (ap) (lf) (PoN) appear 48.1.1 (p)
3 Adjectives large great (V+) difficult	4 Adjectives dry (AntPoV) slow global (LF) young	3 Adjectives wet (Ant) terrible (V-) familiar	4 Adjectives rotten (LF) responsible grey larger
2 Adverbs seriously farther (further) (Cmp)	2 Adverbs far mostly	2 Adverbs eventually most seriously (Sup)	2 Adverbs safely partially

Key
A / C – Abstract / Collective (concrete not labeled)
LF – Low Frequency
L / N – Living / Non-Living
R – Randomly selected
B / N – Biological / Non-biological
PoV / PoN / PoA – Polysemous Verb / Polysemous Noun / Polysemous adjective pairing
NB WORDS IN RED must not be moved to other columns without checking side effects

Appendix 3.2 Randomised word blocks for experiment 3.1

Block A1 Randomised	Block A1 Reversed	Block A2 Randomised	Block A2 Reversed
Nouns	*Nouns*	*Nouns*	*Nouns*
Heap	Variety	Diamond	Boat
Ear	Key	Lemon	Coffee
Cat	Conflict	Delay	Parent
Weapon	Weapon	Pleasure	Pleasure
Conflict	Cat	Parent	Delay
Key	Ear	Coffee	Lemon
Variety	Heap	Boat	Diamond
Verbs	*Verbs*	*Verbs*	*Verbs*
Relax	Laugh	Punish	Change
Hope	Climb	Drink	Drink
Climb	Hope	Change	Punish
Laugh	Relax	*Adjectives*	*Adjectives*
Adjectives	*Adjectives*	Slow	Young
Large	Difficult	Dry	Global
Great	Great	Global	Dry
Difficult	Large	Young	Slow
Adverbs	*Adverbs*	*Adverbs*	*Adverbs*
Farther / further	Seriously	Mostly	Far
Seriously	Farther / further	Far	Mostly

Block B1 Randomised	Block B1 Reversed	Block B2 Randomised	Block B2 Reversed
Nouns	*Nouns*	*Nouns*	*Nouns*
Doubt	Hill	Hen	Prejudice
Fire	Nation	Guitar	Quay
Fly	Rose	Crowd	Shoe
Chair	Chair	Temperature	Temperature
Rose	Fly	Shoe	Crowd
Nation	Fire	Quay	Guitar
Hill	Doubt	Prejudice	Hen
Verbs	*Verbs*	*Verbs*	*Verbs*
Eat	Bargain	Appear	Dry
Run	Cover	Delay	Delay
Cover	Run	Dry	Appear
Bargain	Eat	*Adjectives*	*Adjectives*
Adjectives	*Adjectives*	Grey	Larger
Terrible	wet	Responsible	Rotten
Familiar	Familiar	Rotten	Responsible
Wet	Terrible	Larger	Grey
Adverbs	*Adverbs*	*Adverbs*	*Adverbs*
Most seriously	Eventually	Partially	Safely
Eventually	Most seriously	Safely	Partially

Appendix 3.3 Experiment 3.1 instructions

Semantic Similarity Study

Thank-you for agreeing to help with this study.

Most of the pages that follow have an instruction (containing a particular word) at the top and 2 boxes on them.

Here is an example of the kind of instruction you will see:

Please print (in BLOCK CAPITALS) **two** sentences, **between 10 and 20 words long** in the boxes below, using the Noun FIGHT

Your sentences should be in the form of either

- A question
- An instruction
- A commitment
- An expression

Or

- A declaration

A separate sheet is available to explain these forms.

If the word seems to have more than one meaning you can choose which one to use but please use the same meaning in each of the two sentences and stick to the type of word given in the instructions. For example, if it is the Noun FIGHT please don't use it as a Verb.

Important
This study is **not** measuring you for intelligence or creativity; it is looking at the everyday use of the English Language.

We want you to help us by writing natural sentences. That is they must be meaningful sentences that you might actually say or write, or that other people might use to communicate with you. Please don't stick to a single form of sentence; they can be statements, instructions, commitments, expressions or declarations.

Please ask an investigator if you have any questions before continuing…

otherwise please turn over

Appendix 3.4 Experiment 3.1 Extracts from sentence capture form

The next 3 pages are about VERBS

A verb, such as *to fight*, is a word that expresses an action or a state of being. You can use any valid form of the Verb in the sentence you write for example:

fight … as in … I, you, we, they … fight
fights … as in … he, she … fights
… fought …
… fighting …

but NOT the Noun sense of the word, for example:

… the fight …
… a vicious fight …
… many fights …

Please ask an investigator if you have any questions before continuing…

otherwise please turn over

Please print (in BLOCK CAPITALS) **two** sentences, **between 10 and 20 words long** in the boxes below, using the Verb APPEAR

Your sentences should be in the form of either

- A question
- An instruction
- A commitment
- An expression

Or

- A declaration

A separate sheet is available to explain these forms.

Sentence 1

Sentence 2

Please print (in BLOCK CAPITALS) **two** sentences, **between 10 and 20 words long** in the boxes below, on the general Topic of

Likes & dislikes, expressing opinions

and using the Noun QUAY

Your sentences should be in the form of either

- A question
- An instruction
- A commitment
- An expression

Or

- A declaration

A separate sheet is available to explain these forms.

Sentence 1

Sentence 2

Appendix 3.5 Personal data sheet(for experiment 3.1)

And finally, a few details about yourself . . .

. These are your name, approximate age and the subject of the degree you are studying. You will also be asked to confirm that you are a native speaker of English (i.e. someone for whom it is their first language, spoken since birth).

Name			
Age (tick)	Under 18 ☐	18-22 ☐	Older than 22 ☐
Degree Title			
I confirm that I am a Native English speaker (it is my first language, spoken since birth)		Sign below:	

Please collect your compensation and sign a receipt.

Appendix 3.6 Themes investigated for experiment 3.1

Theme	KS3G	TYE	TYR	NYT
The School (subjects, timetable preferences etc.)	Y	N	N	N
Greeting people and talking about yourself, taking leave	Y	N	Y	Y
Likes & dislikes, expressing opinions	Y	N	Y	Y
Activities, pastimes, sports, clubs, talents	Y	N	N	Y
Shopping & money	Y	Y	Y	Y
Where are you from, background	Y	N	N	Y
About the house (location & movement), household	Y	Y	N	Y
Describing family, relatives, people	Y	Y	Y	Y
Eating out (restaurant etc.)	Y	Y	Y	Y
Directions, tourism, in the street, around the town, where you are going	Y	N	Y	Y
Daily routines (& telling time)	Y	N	Y	Y
Describing people	Y	Y	Y	Y
Travelling and getting about, public transport	Y	N	Y	Y
Dealing with money (transactions, earning, shopping	Y	N	Y	Y
Pets, animals	Y	Y	N	N
Clothes	Y	Y	N	N
Holidays, Hotels, Airports	Y	Y	Y	Y
Shopping for food & drink, drink, ordering drinks	Y	Y	Y	Y
Work & Jobs	Y	Y	N	Y
The Weather	Y	Y	Y	N
Future Events, speaking about th future	Y	N	N	Y
Past Events, saying what you did	Y	N	N	Y
Body & Health, doctor's surgery, feeling ill, injuries	Y	Y	Y	Y
Using the car	Y	N	N	N
Writing a letter	Y	N	N	N
Music (instruments, concert, pop)	Y	N	N	N
Getting help in an emergency	Y	N	N	N
Going out (socially), giving invitations	Y	N	N	Y
Talking about magazines	Y	N	N	N
Complaints & problems	N	N	Y	N
At the post office	N	N	Y	N
Discussing languages	N	N	N	Y
Giving orders and instructions	N	N	N	Y
Giving invitations	N	N	N	Y
Asking permission, for favours	N	N	N	Y
Exchanging news	N	N	N	Y
Living and working abroad	N	N	N	Y

KS3G = Key Stage 3 German, the target attainment 4 of key stage 3 is concerned with the production of texts containing 2 or 3 short sentences on familair topics [Hawkin, T 1995]
TYE = Teach Yourself English [Hunt 2001]
TYR = Teach Yourself Russian [Farmer 1996]
NYT = Now Your Talking (a multimedia Irish course for beginners) [O'Donaill 1995]

Appendix 3.7 Dialogue Act clarification sheet (for experiment 3.1)

Statement:
Statements, descriptions, classifications, explanations

e.g.
Siamese cats are very rare in the part of town that I live in.

Instruction:
Instructions, orders, commands, requests

e.g.
Turn off all off the computers and the printer before you leave the lab tonight.

Commitments:
Promises, vows, pledges, contracts, guarantees

e.g.
I will pay you everything I owe by the end of the month.

Expressions
Apologies, thanks, congratulations, welcomes, condolences

e.g.
I am sorry about damaging your car in the car park last week.

Declarations

Declarations, pronouncements

e.g.

After counting the votes, Adam Taylor is the duly elected representative for Software Engineering.

Appendix 3.8 example Word and Theme Supplement Experiment 3.1

Block A1 Nouns	Block A2 Nouns	Block B1 Nouns	Block B2 Nouns
	Hand HF (Ear A1) 12 Holiday time & travel 12.5 2E Services **consulting a doctor, dentist or chemist**	**Eye** HF (Ear A1) 12 Holiday time & travel 12.5 2E Services **consulting a doctor, dentist or chemist**	**Nose** HF (Ear A1) 12 Holiday time & travel 12.5 2E Services **consulting a doctor, dentist or chemist**
Desk HF (B1 Chair)			
Block A1 Verbs	Block A2 Verbs	Block B1 Verbs	Block B2 Verbs
	31.1 **console VLF** (Relax A1) 14 Theme 4. The Young Person in Society 14.1 4A Character and Personal Relationships **the qualities of a good friend or relationship**	31.1 **calm LF** (Relax A1) 14 Theme 4. The Young Person in Society 14.1 4A Character and Personal Relationships **the qualities of a good friend or relationship**	31.1 **comfort LF** (Relax A1) 14 Theme 4. The Young Person in Society 14.1 4A Character and Personal Relationships **the qualities of a good friend or relationship**
48.1.2 **present LF** (Appear B2) 14 Theme 4. The Young Person in Society 14.1 4A Character and Personal Relationships the **qualities of a good friend or relationship**			
Block A1 Adjectives	Block A2 Adjectives	Block B1Adjectives	Block B2 Adjectives
Green HF (Grey B2)	**White** HF (Grey B2)	**Yellow** LF (Grey B2)	
			Stale VLF (Rotten B2)
Block A1 Adverbs	Block A2 Adverbs	Block B1 Adverbs	Block B2 Adverbs
Already HF (Eventually B1)	**Finally** HF (Eventually B1)		**Still** HF (Eventually B1)
		Across LF (A2 far)	

Appendix 3.9 A sample of sentences captured by Experiment 3.1

Participant Number	Sentence
29	Would you like to drink this wine with your meal?
16	I often like to buy a bottle of Coke before class to drink later.
29	Thank you very much for the offer, but I won't be drinking tonight.
16	Drinking down the pub with mates is a regular activity for me.
6	I am so thirsty, I really need to drink some water before I die of thirst.
13	We stayed up all night drinking wine and chatting.
10	Last Friday me and my friends went drinking in Liverpool.
6	Would you like to go out to drink with me tonight?
13	You can drink your tea but then you've got to go to bed.
10	After I returned from the gym I drank two pints of cold water.
22	Drinking a beer in the Summer is the best way to cool down.
2	I drink about a litre of orange juice a day, because it makes me feel healthy.
22	My cat always seems to be drinking from the plughole.
2	When young people drink alcohol, they are often criticised for being irresponsible or out of control.
3	I am so hungry I could eat a whole horse plus dessert.
17	Sometimes eating food that's bad for you makes you feel better.
14	I love eating chocolate, it's better than any food in the world for taste.
7	I eat chips normally but tonight I fancy a change.
7	I really don't know what to eat tonight so I might go out somewhere.
14	Don't be stupid she'll eat you alive if you go in there.
26	I had to eat all my food, my mother never lets me get away from the table.
17	Eating chocolate always makes everyone feel happier.
3	My friends and I are eating at that new restaurant tonight.
25	I could have eaten another meal, I'm still starving.
25	I would consider eating as a hobby, I love food.
11	I ate some amazing vegan curry last night it was very tasty.
11	I always have to be careful about what I eat, in case it contains animal products.
26	He had a very good meal today, he ate some fish with chips.

Appendix 3.10 Sentences selected for paraphrase generation

Stimulus Word / DA	Sentence	SP
Laugh Commit	I will make you laugh so much that your sides ache.	71
Parent Express	I offer my condolences to the parents of John Smith, who was unfortunately murdered.	79
Young Commit	If you continuously use these products, I guarantee you will look very young.	81
Safely State	We got home safely in the end, although it was a long journey.	88
Far/farther Quest	Will I have to drive far to get to the nearest petrol station?	112
Hill Instruct	Meet me on the hill behind the church in half an hour.	107
Drink Quest	Would you like to drink this wine with your meal?	120
Pleasure Declare	It gives me great pleasure to announce the winner of this year's beauty pageant.	110
Heap Express	I am sorry but I can't go out as I have a heap of work to do.	114
Wet Instruct	Get that wet dog off my brand new white sofa.	116
Climb Quest	Could you climb up the tree and save my cat from jumping please?	123

Appendix 3.11 A sample paraphrased sentence capture sheet (experiment 3.1 modifications to materials)

Thank-you for agreeing to help with this study. You can withdraw from the study if you wish at any time before returning this questionnaire.

This study is not testing you in any way, it is to produce data that can be used in future experiments which measure the similarity of sentence pairs.

We would like you to paraphrase 3 sentences. By paraphrase we mean we want you to express the same sentence in a different way.

In each case the sentence you write should come as close as you can manage to meaning the same thing as the example supplied to you

The sentences should be between 10 and 20 words long

You will be told that ther is one word that you MUST use in the sentence you write

Otherwise please try not to use more words form the original sentence than you have to

Please PRINT the sentences in the boxes provided as they will have to by types into a computer

1. Your sentence must contain the verb LAUGH

Example	I will make you laugh so much that your sides ache.
Your version	

2. Your sentence must contain the adverb SAFELY

Example	We got home safely in the end, although it was a long journey.
Your version	

3. Your sentence must contain the noun PARENTS

Example	I offer my condolences to the parents of John Smith, who was unfortunately murdered.
Your version	

Appendix 3.12 A sample of paraphrased sentences for high similarity pairs (experiment 3.1 modifications to materials)

Original sentence: **If you continuously use these products, I guarantee you will look very young.**

Alternatives:
I can promise you that so long as you use these products without interruption you will appear to be young.
If you use these things all the time, I promise they'll keep your appearance young.
You will look young if you use this product continuously.
In order to keep feeling young I would advocate an active and healthy lifestyle.
Trust me, if you use these cosmetics every day they will keep you looking young.

Selected:
I assure you that, by using these products consistently over a long period of time, you will appear really young.

Original sentence: **Will I have to drive far to get to the nearest petrol station?**

Alternatives:
Is it much farther to get by car to a filling station?
How much farther is it to the next petrol station.
Must I convey this motorised vehicle much farther to reach the closest automobile service point?
Children always say "do we have to go any farther?", when you take them on a long walk.
Have I got to travel much farther to reach the next petrol station.

Selected:
Is it much farther for me to drive to the next gas station?

Original sentence: **Meet me on the hill behind the church in half an hour.**

Alternatives:
Be on the hill to meet to me at the back of the Church in thirty minutes.
We will meet at the church behind the hill in half an hour.
Rendezvous with me on the hll at the back of the house of God in thirty minutes.
Hill walking is wonderful because it gives you the chance to experience great scenery.
I will be at the top of the hill on the other side of the church in half an hour, meet me there.

Selected:
Join me on the hill at the back of the church in thirty minutes time.

Appendix 3.13 Experiment 3.2 Selected sentence pairs with predicted similarity bands and selection criteria

Stim words = stimulus word for each sentence, followed by theme, below, where used. SP column = sentence pair no. on cards, with predicted similarity band below. Sentence = each sentence in the pair. Criteria = language representation criteria. The ID column refers to primary key in sentence database. Target high similarity pairs are shown in Red.

High – 15 pairs

Stim Words	SP	ID	Sentence	Criteria
Seriously Most seriously	67	138	I advise you to treat this matter very seriously as it is vital.	Adverb-Adverb Superlative Instruction/Instruction
	H	955	You must take this most seriously, it will affect you.	
laugh	71	1024	I will make you laugh so much that your sides ache.	To meet high similarity target Paraphrased sentence (SP71) Commitment/Commitment
	H		When I tell you this you will split your sides laughing.	
Hill-Quay Going_out	73	184	Do you want to come with us to the pub behind the hill?	Concrete Noun-Noun To meet high similarity target Same theme / different word Question/Statement
	H	423	We are going out for drinks tonight in Salford Quays if you would like to come.	
parent	79	1088	I offer my condolences to the parents of John Smith, who was unfortunately murdered.	Paraphrased sentence (SP79) Expression/Expression
	H		I express my sympathy to John Smith's parents following his murder.	
Lemon	83	581	I always like to have a slice of lemon in my drink especially if it's Coke.	To meet high similarity target Noun Low Frequency Living Biological
	H	141	I like to put a wedge of lemon in my drinks, especially cola.	
Diamond	89	276	A man called Dave gave his fiancee a large diamond ring for their engagement.	Noun Non-Living/Non-Biological To meet high similarity target Low Frequency
	H	451	The man presented a diamond to the woman and asked her to marry him.	

Global	97	471	Global warming is what everyone is worrying about today.	To meet high similarity target Low Frequency Adjective
	H	381	The problem of global warming is a concern to every country in the world at the moment.	
Terrible	101	507	There was a terrible accident, a pileup, on the M16 today.	To meet high similarity target Adjective Value negative
	H	940	It was a terrible accident, no one believed it was possible.	
Coffee	103	575	The first thing I do in a morning is make myself a cup of coffee	To meet high similarity target Non-Living/Biological
	H	817	The first thing I do in the morning is have a cup of coffee.	
Run	127	194	We often ran to school because we were always late.	To meet high similarity target Verb
	H	986	I knew I was late for my class so I ran all the way to school.	
Bargain	109	80	We tried to bargain with him but it made no difference, he still didn't change his mind.	To meet high similarity target Verb
	H	952	I tried bargaining with him, but he just wouldn't listen.	
Familiar	113	76	I think I know her from somewhere because she has a familiar face.	Adjective To meet high similarity target Statement/Question
	H	942	You have a very familiar face, where do I know you from?	
Eat	131	192	I am so hungry I could eat a whole horse plus dessert.	Verb To meet high similarity target
	H	946	I could have eaten another meal, I'm still starving.	
Punish	70	1123	You must realise that you will definitely be punished if you play with the alarm.	Verb To meet high similarity target
	H	1124	He will be harshly punished for setting the fire alarm off.	
Rose	121		Roses can be different colours, it has to be said red is the best though.	Noun To meet high similarity target
	H	1125	Roses come in many varieties and colours , but yellow is my favourite.	

High-Medium - 5

Present Good_friend	76	1038	You're not a good friend if you're not prepared to be present when I need you.	To meet similarity target Verb Same word / same theme
	H-M	764	A good friend always seems to be present when you need them.	
variety	78	62	I have invited a variety of people to my party so it should be interesting.	To meet high similarity target

	H-M	741	A number of invitations were given out to a variety of people inviting them down the pub.	Adjective Same word / same theme
Farther	82	757	We ran farther than the other children that day.	To meet similarity target Adverb Same word
	H-M	563	You ran farther than anyone today.	
Ear Doctor / dentist	84	603	It seems like I've got eczema on my ear doctor, can you recommend something for me?	To meet similarity target Noun Hand (derived from Ear) Same word (SP84) Question/Statement
	H-M	485	I had to go to a chemist for a special rash cream for my ear.	
Safely	88	677	We got home safely in the end, although it was a long journey.	Safely To meet similarity target Paraphrased sentence. (SP88)
			Though it took many hours travel, we finally reached our house safely.	

Medium-High 5

Large Larger	94	250	Peter was a very large youth, whose size intimidated most people, much to his delight.	Adjective-Adjective Comparative
	M-H	789	Now I wouldn't say he was fat, but I'd certainly say he was one of the larger boys.	
Dry Wet	96	672	We got soaked in the rain today, but now we are nice and dry.	Adjective Antonyms
	M-H	334	I was absolutely soaking wet last night, I drove my bike through the worst weather.	
Eventually Safely	102	522	After hours of getting lost we eventually arrived at the hotel.	High Frequency Adverbs
	M-H	39	After walking against the strong wind for hours he finally returned home safely.	
Difficult Mostly	106	396	It is mostly very difficult to gain full marks in today's exam.	To meet similarity target Adjective / Adverb Same word
	M-H	1017	The exam was really difficult, I've got no idea if I'm going to pass.	
Console	108	1117	If you don't console with a friend, there is a chance you may hurt their feelings.	To meet similarity target Verb Relax Levin class 3.1 Same word / same theme Good
	M-	605	One of the qualities of a good friend is the ability to console.	

	H			Friend

Medium - 10

Far Farther	112	395	Will I have to drive far to get to the nearest petrol station? Is it much farther for me to drive to the next gas station?	Adverb Place Comparative Paraphrased sentence derived from far but using Farther instead. (SP112) Question/Question
Cat Hen	118 M	7 1043	The cat was hungry so he went into the back garden to find lunch. The hen walked about in the yard eating tasty grain.	Living Biological Nouns
Pleasure Doubt	124 M	144 976	The pleasure that I get from studying, is that I learn new things. I have a doubt about this exam, we never got to study for it.	Abstract emotion Positive vs Negative
Mostly Partially	126 M	600 1072	The damp was mostly in the very corner of the room. The young lady was somewhat partially burnt from the sun.	Two Low Frequency Adverbs Frequency / Degree
Drink Eat	66 M	389 86	Would you like to go out to drink with me tonight? I really don't know what to eat tonight so I might go out somewhere.	Levin Class 39.1 Question/Statement
Cover	72 M	631 950	You shouldn't be covering what you really feel. There is no point in covering up what you said, we all know.	State verb
Appear	75 M	434 1066	The ghost appeared from nowhere and frightened the old man. The ghost of Queen Victoria appears to me every night, I don't know why, I don't even like the royals.	Process verb
Young	81	1098	If you continuously use these products, I guarantee you will look very young. I assure you that, by using these products consistently over a long period of time, you will appear really young.	Adjective Commitment Paraphrased sentence (SP81)
Temperature	87	1050	Water freezes at a certain temperature, which is zero degrees Celsius.	Noun: to allow word Temperature to participate

	M	210	The temperature of boiling water is 100 C and the temperature of ice is 0 C.	9
Responsible	92	118	Because I am the eldest one I should be more responsible.	Noun
	M	866	Just because of my age, people shouldn't think I'm a responsible adult, but they do?	To meet similarity target Contains a human error - 866 is not a question

Medium-Low - 5

Nation	85	937	I am proud of our nation, well, most of it.	Abstract Noun
	M-L	609	I think of myself as being part of a nation.	
Fire	93	68	I need to dash into the kitchen because I think my chip pan is on fire.	To meet similarity target Noun Same word Statement/Instruction
	M-L	495	In the event of a chip pan fire follow the instructions on the safety note.	
Grey	95	217	I'm going to buy a grey jumper today, in half an hour.	To meet similarity target Adjective Same word Statement/Question
	M-L	670	That's a nice grey top, where did you get it from?	
Delay	681	288	When I was going out to meet my friends there was a delay at the train station.	Polysemous Noun/Verb combination
	M-L	1068	The train operator announced to the passengers that the train would be delayed.	
Mostly Most seriously	105	639	I'm worried most seriously about the presentation, not the essay.	Adverbs HF / LF and near homograph
	L-M	396	It is mostly very difficult to gain full marks in today's exam.	

Low-Medium - 5

Great	100	706	That's not a very good car, on the other hand mine is great.	Adjective Value + / -

Terrible	L-M	186	This is a terrible noise level for a new car.	
Seriously Difficult	128	30	I hope you're taking this seriously, if not you can get out of here.	Adverb Manner / Adjective Difficult Indirect Instruction / Statement
	L-M	748	The difficult course meant that only the strong would survive.	
Conflict Prejudice	117	93	He fought in the war in Iraq before being killed in a car crash.	Abstract State / Emotion negative affect
	L-M	780	The prejudice I suffered whilst on holiday in Iraq was quite alarming.	
Lemon Rose	119	142	My bedroom wall is lemon coloured but my mother says it is yellow.	Living Biological Fruits&Vegs vs Plants&Flowers
	L-M	325	Roses can be different colours, it has to be said red is the best though.	
Rotten Global	125	430	The perpetrators of war crimes are rotten to the core.	Two low frequency Adjectives
	L-M	158	There are many global issues that everybody should be aware of, such as the threat of terrorism.	

Low - 15

Guitar Weapon	130	657	I bought a new guitar today, do you like it?	Nouns Miscellaneous Artefact / Musical Instrument Question/Statement
	L	537	The weapon choice reflects the personality of the carrier.	
Relax Dry	69	1030	Does music help you to relax, or does it distract you too much?	Two State ProcessVerbs (error – Dry adjective) Question/Question
	L	1102	Does this sponge look wet or dry to you?	
Key Quay	74	2	This key doesn't seem to be working, could you give me another?	Homophone pair Question/Statement
	L	860	I dislike the word quay, it confuses me, I always think of things for locks, there's another one.	
Boat Chair	80	150	Boats come in all shapes and sizes but they all do the same thing.	Normally found indoors/outdoors
	L	322	Chairs can be comfy and not comfy, depending on the chair.	
Heap crowd	86	12	There was a heap of rubble left by the builders outside my house this morning.	Collective Nouns living/non-living
	VL	106	Sometimes in a large crowd accidents may happen, which can cause deadly injuries.	

Run Climb	90 L	338 554	I used to run quite a lot, in fact once I ran for North Tyneside. I used to climb lots at school as we had a new climbing wall put in the gym.	Two Action Verbs Levin Class 51.3.2
Laugh Bargain	91 VL	553 198	I love to laugh as it makes me happy as well as those around me. I thought we bargained that it would only cost me a pound.	Two Action Verbs, one LF
Punish Delay	98 L	474 794	He was harshly punished for setting the fire alarms off. He delayed his response, in order to create a tense atmosphere.	Two Action-Process verbs
Change Appear	104 VL	301 675	Someone spilt a drink accidentally on my shirt, so I changed it. It appears to have shrunk, it wasn't that size before I washed it.	Two Process Verbs
Safely Seriously	77 VL	1073 266	The children crossed the road very safely thanks to the help of the lollipop lady. It was feared that the child might not recover, because he was seriously ill.	Two Manner Adverbs
Responsible Slow	115 L	1060	The responsible man felt very guilty when he crashed into the back of someone's car. A slow driver can be annoying even though they are driving safely.	Two adjectives, to allow these words to participate
Fly Dry (verb)	122 L	321 229	Flies can also carry a lot of disease and cause maggots. I dry my hair after I wash it or I will get ill.	Noun-Verb, to allow these words to participate
Hill	107	938	Meet me on the hill behind the church in half an hour. Join me on the hill at the back of the church in thirty minutes time.	Noun Hill Paraphrased sentence. (SP107) Instruction/Instruction
Hope cover	111 L	258 340	They said they were hoping to go to America on holiday. I like to cover myself up in lots of layers, I don't like the cold.	Hope-cover Two state verbs
Drink	120	477	Would you like to drink this wine with your meal? Will you drink a glass of wine while you eat?	Drink Paraphrased sentence (SP120) Question/Question

Speech Acts test – permutations on 603 and 485

pleasure	110	456	It gives me great pleasure to announce the winner of this year's beauty pageant. It's a real pleasure to tell you who has won our annual beauty parade.	Paraphrased sentence (SP110) Declaration/Declaration

heap	114	50	I am sorry but I can't go out as I have a heap of work to do.	(SP114) Expression/Statement
			I've a heap of things to finish so I can't go out I'm afraid.	
wet	116	189	Get that wet dog off my brand new white sofa.	(SP116) Instruction/Instruction
			Make that wet hound get off my white couch – I only just bought it.	
climb	123	24	Could you climb up the tree and save my cat from jumping please?	(SP123) Question/Question
			Can you get up that tree and rescue my cat otherwise it might jump?	

Appendix 3.14 Calculation of Z-scores for discriminability

SDT is concerned with detecting a signal in the presence of noise (Egan, 1975) and has its origins in operational research in the training of radar operators in World War 2. Given a requirement to make a decision on whether or not a signal occurs in the presence of noise there are four possibilities Hit (signal correctly classified as present), Miss (signal incorrectly classified as absent), False Alarm (missing signal incorrectly classified as present) and Correct Rejection (missing signal correctly classified as absent).
A criterion is used by a human to classify the presence or absence of the signal. Shifting the criterion in one direction can increase the number of Hits, at the expense of increasing the number of False Alarms. Shifting it in the other direction will increase the number of Correct Rejections, at the expense of increasing the number of Misses. SDT is concerned with finding the inherent discriminability of the signal and noise, rather than where the criterion should be set.

The measure of discriminability, d', is calculated using the Z-scores equation A1:

$$d' = z(H) - z(FA)) \qquad (A1)$$

where Z(H) is the Z-score of the hit rate and Z(FA) is the Z-score of the false alarm rate. The larger the value of d' the more discriminable the signal is from the noise.

This can be illustrated with an example, shown in table 3.16. Suppose an event is classified into one of two categories. Class 1 classified 90% accurately; Class 2 classified 70% accurately.

	Response		
Stimulus	Class 1	Class 2	Sum
Class 1	90 *(Hit)*	30 *(Miss)*	120
Class 2	10 *(False Alarm)*	70 *(Correct Reject)*	80
Sum	100	100	200

Table 3.16 Contingency table for detecting Class 1 in the presence of Class 2

The first step is to build a contingency table (table 3.16). The next step is to calculate values known as Z-scores for the hits and false alarms:

	Z-scores
p(hit) = 0.75	0.67449
p(FA) = 0.25	-1.150349

d' = 1.82484

(note that the same discriminability is calculate for detecting Class 2 in the presence of Class 1)

The Z-score is the standard score, and is calculated in statistics using equation A2:

$$z = \frac{X - \overline{X}}{s} \qquad \text{(A2)}$$

Where X is the raw score, $\overline{X}$ is an estimation of the population mean and s is an estimation of the population standard deviation.
Where probabilities are known, as in the examples above it can be calculated from a normal distribution table. This can also be done using the function NORMSINV in Microsoft excel.

Appendix 3.15 Experiment 3.3 extracts from pararagraph capture form

Semantic Similarity Study

Thank-you for agreeing to take part in this study. You may still withdraw before starting the questionnaire or at any point before completing it.

In this experiment we would like you to help us by writing some paragraphs each of which contains a particular sentence that we will supply.

- Each paragraph should be between **three** and **five** sentences long using the sentence we supply.
- You must use this sentence **exactly as we provide it** with no modifications.
- The sentence can be in the middle of the paragraph or at the beginning or end, but it must make sense. That is, the paragraph and the sentence must be about the same subject.
- Please print your paragraphs in BLOCK CAPITALS to make it easier for us to type them into the computer.

This study does not evaluate you in any way; you are providing us with data that we can use to investigate the properties of natural English sentences. So there are no "right" or "wrong" answers, except in the sense that the right answer to each question is one that is natural for you.

Above all, we are looking for natural paragraphs. Please write the kind of paragraph that you would be likely to use in a real conversation, a phone call, a letter, e-mail or Internet forum. Alternatively, try to think of a paragraph that someone might use with you in one of these circumstances. If all else fails try to think of a paragraph that someone would use in a magazine, radio or television programme.

An example follows on the next page.

Please start at the first page and work through the pages in order from start to end. Please take your time over section A; it is better for us to get paragraphs that you have thought about carefully. However, if you have time left over please carry on and do some of section B until you run out of time.

If you have any problems, questions or comments please speak to one of the investigators

. . . Otherwise please turn over and begin.

SP93
Please print (in BLOCK CAPITALS) a paragraph, between 3 and 5 sentences long in the boxes below, using the sentence:

I need to dash into the kitchen because I think my chip pan is on fire.

Paragraph

N.B. Each form contains 8 stimulus sentences. All 8 appear in section A, followed by a repetition of all 8 (in the same order) in section B. For 8 sentence pairs there are 16 sentences. This accommodated by putting one of each pair in Block C and the other in Block D (avoids spurious semantic overlap). It was assumed (correctly) that not all participants would be able to generate 16 paragraphs. However, having each block in a Forward and Reversed order promotes the likelihood of getting an even distribution of paragraphs generated by the sentences.

Appendix 3.16 A sample of the paragraphs selected for the blocked design experiment

SP93_C		*I need to dash into the kitchen because I think my chip pan is on fire.*
Participant	P	Paragraph
1	1	You'll have to hang on for five minutes, I'm afraid. <---------- *insert sentence here* ----------> There's a smell of smoke. I'll get back to you later.
	2	Can you smell something from the other room? Just hold that for a minute will you until I go and see. <---------- *insert sentence here* ---------->
5	1	<---------- *insert sentence here* ----------> It may take me sometime to deal with this. But I will call you back when I can. Must go. Bye for now.
	2	Why do I need to dash now? <---------- *insert sentence here* ----------> Believe it or don't believe it. But I must go now.
6	1	Sorry, sorry, must go. <---------- *insert sentence here* ----------> Call you later.
	2	So there I am, chatting away. Then, oh my god! <---------- *insert sentence here* ----------> Luckily I hadn't even turned the cooker on.
7	1	I awake suddenly. <---------- *insert sentence here* ----------> Then I remember where I am – in a hotel. No kitchen, no chip pan.
	2	<---------- *insert sentence here* ----------> Quick, give us a hand. That's the last time I make chips in this house.
8	1	I can't stop and talk now. <---------- *insert sentence here* ----------> Speak to you later.
	2	When the Jehovah's Witnesses call I have a string of excuses ready. This one is for when it gets really dire. <---------- *insert sentence here* ->
11	1	Sorry to have to cut you short. <---------- *insert sentence here* ----------> I'll ring you back later.
	2	<---------- *insert sentence here* ----------> That's what I told her. It was a ridiculous lie but the only thing I could think of to get away from phone. She is such a terrible bore.
12	1	Oh, good heavens! I have to ring off now. <---------- *insert sentence here* ----------> I'll call you back later, when I've dealt with it.
	2	I don't like the look of that smoke. <---------- *insert sentence here* ----------> I completely forgot that I'd left it on the hob.

Appendix 3. 17 Extra information for unsupervised participants (experiment 3.4)

5, Clifton Rd.,
Chorlton
Manchester
M21 8UX
13/06/08

Extra information for unsupervised participants

If you are completing this by post, please check your questionnaire for any responses that you might have missed before returning it. Your compensation will be sent by return of post when I receive the completed questionnaire.

The questionnaire contains 4 different sections, each of which uses a particular pair of sentences. Each judgement involves deciding which sentence in the pair belongs to a particular paragraph.

You may find that you are quicker in finishing some of the questions than others. You may also find that within a section you are quicker in making some of the judgements than others. This is perfectly normal, but the judgements and sections have been chosen to make the effort required from each of the participants approximately equal.

There are right and wrong answers to the questions, but I need to point out that right or wrong, your answers are equally valuable provided that they are you are your best judgements based on the **meanings** of the sentences. What I am trying to discover is how easily one sentence can be substituted (or mistaken) for another.

So please don't use any other method to try and work out the correct answers – for example you can not assume there are equal numbers of A and B answers for a given pair of sentences, they have been randomised in a way that prevents this.

This work should take you between an hour and one-and-a-half hours to complete based on timings form a pilot study. Please arrange to have enough time, undisturbed to complete the work in one session. Please go through it from start to end and don't go back to revise answers. If you need to take a break there is a natural point half way through – you will find I because the cover sheet appears again there.

Thanks you for helping. It is very important for me to get all of the questionnaires back – your work is designed to be compiled with other participants and (because of the randomisation etc.) if I don't get it back it will be very difficult generate a replacement. Once gain, please check it for missed answers before you return it.

Thanks,

Jim O'Shea

j.d.oshea@mmu.ac.uk

0161 247 1546

Appendix 3.18 Block design for Experiment 3.4

Expt Block structure	Phase 1							
SP120-124F					SP124-120R			
1	2	3	4		5	6	7	8
SP120-124F	SP107-126R	SP126-107F	SP120-124F		SP124-120R	SP126-107R	SP107-126F	SP124-120R
SP107-126F	SP120-124F	SP120-124F	SP126-107R		SP126-107F	SP124-120R	SP124-120R	SP107-126R
Student	Non-student	Student	Non-student		Student	Non-student	Student	Non-student
SP124-120F					SP120-124R			
9	10	11	12		13	14	15	16
SP124-120F	SP107-126R	SP126-107F	SP124-120F		SP120-124R	SP107-126R	SP126-107F	SP120-124R
SP107-126F	SP124-120F	SP124-120F	SP126-107R		SP107-126F	SP120-124R	SP120-124R	SP126-107R
Student	Non-student	Student	Non-student		Student	Non-student	Student	Non-student
SP88-95F					SP95-88R			
17	18	19	20		21	22	23	24
SP88-95F	SP72-93R	SP93-72F	SP88-95F		SP95-88R	SP72-93R	SP93-72F	SP95-88R
SP72-93F	SP88-95F	SP88-95F	SP93-72R		SP72-93F	SP95-88R	SP95-88R	SP93-72R
Non-student	Student	Non-student	Student		Non-student	Student	Non-student	Student
SP95-88F					SP88-95R			
25	26	27	28		29	30	31	32
SP95-88F	SP72-93R	SP93-72F	SP95-88F		SP88-95R	SP72-93R	SP93-72F	SP88-95R
SP72-93F	SP95-88F	SP95-88F	SP93-72R		SP72-93F	SP88-95R	SP88-95R	SP93-72R
Non-student	Student	Non-student	Student		Non-student	Student	Non-student	Student

Expt Block structure Phase 2

SP120-124F			
33	34	35	36
SP120-124F	SP107-126R	SP126-107F	SP120-124F
SP107-126F	SP120-124F	SP120-124F	SP126-107R
Non-student	Student	Non-student	Student

SP124-120R			
37	38	39	40
SP124-120R	SP126-107R	SP107-126F	SP124-120R
SP126-107F	SP124-120R	SP124-120R	SP107-126R
Non-student	Student	Non-student	Student

SP124-120F			
41	42	43	44
SP124-120F	SP107-126R	SP126-107F	SP124-120F
SP107-126F	SP124-120F	SP124-120F	SP126-107R
Non-student	Student	Non-student	Student

SP120-124R			
45	46	47	48
SP120-124R	SP107-126R	SP126-107F	SP120-124R
SP107-126F	SP120-124R	SP120-124R	SP126-107R
Non-student	Student	Non-student	Student

SP88-95F			
49	50	51	52
SP88-95F	SP72-93R	SP93-72F	SP88-95F
SP72-93F	SP88-95F	SP88-95F	SP93-72R
Student	Non-student	Student	Non-student

SP95-88R			
53	54	55	56
SP95-88R	SP72-93R	SP93-72F	SP95-88R
SP72-93F	SP95-88R	SP95-88R	SP93-72R
Student	Non-student	Student	Non-student

SP95-88F			
57	58	59	60
SP95-88F	SP72-93R	SP93-72F	SP95-88F
SP72-93F	SP95-88F	SP95-88F	SP93-72R
Student	Non-student	Student	Non-student

SP88-95R			
61	62	63	64
SP88-95R	SP72-93R	SP93-72F	SP88-95R
SP72-93F	SP88-95R	SP88-95R	SP93-72R
Student	Non-student	Student	Non-student

Appendix 3.19 Instructions for Experiment 3.4 (discriminability rating)

Thank-you for volunteering to take part in this study

This study does not test you in any way, we are testing the properties of natural English sentences. There are a few personal questions at the end just to make sure we get a good variety of the people in the experiment.

You do need to be a Native English speaker to help with this study. That is English should be your main language and the first language you learnt. If you aren't a native English speaker please return the questionnaire without filling it in.

We would like you to help us by matching some sentences to paragraphs.

Each page in the questionnaire contains a paragraph with a missing sentence.

There is a paragraph at the top of the page and there are two sentences (A and B) below it.

After the paragraph was written, we removed one of these two sentences from it and replaced with the text <---------- *insert sentence here* ---------->

Please read the paragraph and both of the sentences carefully, then decide which sentence belongs in the paragraph.

Tick the box at the bottom of the page to show which sentence you have chosen.

Please start at the first page and work through in the pages in order from start to finish.

There is an example at the start.

If you have any problems, questions or comments please speak to one of the investigators

. . . Otherwise please turn over and begin.

Appendix 3.20 Extract from discriminability rating form

EXAMPLE

Please read the paragraph and the following sentences carefully:

Paragraph
I am feeling very annoyed today. <---------- *insert sentence here* ----------> I know I have to look after them but they are old enough to do some things for themselves.

Sentences	
A	The kids have left a heap of their clothes on the floor again.
B	You should not be allowed to see that film at your age.

Please tick the box to show which sentence belongs in the paragraph:

Sentence A ☑

Sentence B ☐

EXAMPLE

Appendix 4.1 A sample of the 600 Straightforward Questions from the dataset

am i able to export my emi to any other applications
are alimony payments considered taxable income
at what times will the lorries deliver
can a husband and wife run a business as a sole proprietor or do they need to be a partnership
do you know where i can get the flu immunisation done privately
how can a man color his moustache and sideburns
if i apply online can i apply again via telephone
is there anything i can do to jazz it up
may i claim the excess expenses on form 8863
must i file electronically to split my refund
should i compliment a company for its products
were not childhood diseases disappearing long before vaccines were introduced
what kind of interest and penalties will i be charged for filing and paying my taxes late
when did tattooing start
where are the enlarged group's headquarters
where do i find a good artist and what should i look for in a tattoo artist
which fats and oils are best suited to frying and deepfrying fish products
who are the composers and performers for the classical guitar?
why are there products on the site that i cannot buy online
will the research lead participants to reveal involvement in illegal activities in their past or present

Appendix 4.2 A sample of the 600 Straightforward Non-questions

you decide
it was a bad time
it is left us all completely shocked
much loathed grocery department manager robert suddenly disappeared in middecember
the people dressed like labtechnicians need to be banned immediately
there is room for improvement here in my opinion
this is an issue on my end not his
we were trying to fix an old bamboo sofa
yes the lc3 still comes standard with electro harmonix tubes
china's president hu jintao began to talk of building a harmonious society
physicians' services may be rated ranked advertised or otherwise promoted online
please do not go flaming the good major for squelching my speech
so the next day mrs snot arrives for her seven large boxes
and you can be very very tired travelling all the time
we are very excited to sponsor and attend the milblog conference
but then i remember that there is still six weeks left
he has an impressive trackandfield record including a mile best of 419
i really fancy a home cinema surround sound system from sony
so touring comes as a result of a good or interesting musical profile
that is one reason why nondoms have found london a happier place to be
the clean geometric shapes and colors agnes and agnes design puts together are fantastic
they can often look quite red for 2 days after using the product

Appendix 4.3 A sample of the 600 Difficult Nonquestions

what great news jim
what we want is watney's
what would be said by the chairman was too unpleasant to think about
when melting ice expands it will probably crack the pipes
when is less important than where
where legal disputes arise lawyers grow fat
where it is on show next i will get there somehow
where the virus struck complete chaos followed
which bike sold best the manufacturers would know
which colour hair suited her best no one could agree on
which year ireland last won the grand slam will be on the net somewhere
who dares wins
who laughs last does not always laugh loudest in my experience
who is the better of the two this contest should decide
why moths developed wing eyemarkings is easily explained by evolution
why instead of negotiating he fired will be determined by the inquiry
why it happened is a mystery at present
how depressing
how swimming tones muscles of even the most sedentary person
how migrating birds navigate or current theory about it makes riveting reading
how steam cleans away even the stiffest grease

A full listing of the 600 straightforward questions, straightforward nonquestions and difficult non-questions is available from the author.

Appendix 4.4 Contraction list lookup table

i'd	i would	when'll	when will	weren't	were not
i'll	i will	when's	when is	won't	will not
i'm	i am	when've	when have	wouldn't	would not
i've	i have	where'd	where would	wasn't	was not
you'd	you would	where'll	where will	might've	might have
you'll	you will	where's	where is	must've	must have
you're	you are	where've	where have	should've	should have
you've	you have	who'd	who would	would've	would have
he'd	he would	who'll	who will	could've	could have
he'll	he will	who's	who is	will've	will have
he's	he is	who've	who have	let's	let us
she'd	she would	why'd	why would	'tis	it is
she'll	she will	why'll	why will	'twas	it was
she's	she is	why's	why is	whatever's	whatever is
it'd	it would	why've	why have	whoever's	whoever is
it'll	it will	how'd	how would	whenever's	whenever is
it'd	it would	how'll	how will	wherever's	wherever is
it's	it is	how's	how is	whyever's	whyever is
we'd	we would	how've	how have		
we'll	we will	here's	here is		
we're	we are	ain't	are not		
we've	we have	aren't	are not		
they'd	they would	can't	can not		
they'll	they will	couldn't	could not		
they're	they are	could've	could have		
they've	they have	daren't	dare not		
there's	there is	didn't	did not		
there'd	there would	doesn't	does not		
there're	there are	don't	do not		
that'll	that will	hadn't	had not		
that's	that is	hasn't	has not		
that've	that have	haven't	have not		
that'd	that would	isn't	is not		
this'll	this will	mayn't	may not		
this'd	this would	might'nt	might not		
what'd	what would	mustn't	must not		
what's	what is	needn't	need not		
what'll	what will	ought'nt	ought not		
what've	what have	shan't	shall not		
when'd	when would	shouldn't	should not		

Appendix 4.5 Function word table used for tokenising sentences

a 1	between 45	hereby 89
about 2	beyond 46	herein 90
above 3	both 47	hereupon 91
across 4	but 48	hers 92
after 5	by 49	herself 93
afterwards 6	can 50	him 94
again 7	cannot 51	himself 95
against 8	could 52	his 96
all 9	despite 53	how 97
almost 10	did 54	however 98
alone 11	do 55	i 99
along 12	does 56	ie 100
already 13	done 57	if 101
also 14	down 58	in 102
although 15	during 59	indeed 103
always 16	each 60	inside 104
am 17	eg 61	instead 105
among 18	either 62	into 106
amongst 19	else 63	is 107
amoungst 20	elsewhere 64	it 108
an 21	enough 65	its 109
and 22	etc 66	itself 110
another 23	even 67	last 111
any 24	ever 68	latter 112
anyhow 25	every 69	latterly 113
anyone 26	everyone 70	least 114
anything 27	everything 71	less 115
anyway 28	everywhere 72	lot 116
anywhere 29	except 73	lots 117
are 30	few 74	many 118
around 31	first 75	may 119
as 32	for 76	me 120
at 33	former 77	meanwhile 121
be 34	formerly 78	might 122
became 35	from 79	mine 123
because 36	further 80	more 124
been 37	had 81	moreover 125
before 38	has 82	most 126
beforehand 39	have 83	mostly 127
behind 40	he 84	much 128
being 41	hence 85	must 129
below 42	her 86	my 130
beside 43	here 87	myself 131
besides 44	hereafter 88	namely 132

near 133
need 134
neither 135
never 136
nevertheless 137
next 138
no 139
nobody 140
none 141
noone 142
nor 143
not 144
nothing 145
now 146
nowhere 147
of 148
off 149
often 150
on 151
once 152
one 153
only 154
onto 155
or 156
other 157
others 158
otherwise 159
ought 160
our 161
ours 162
ourselves 163
out 164
outside 165
over 166
per 167
perhaps 168
rather 169
re 170
same 171
second 172
several 173
shall 174
she 175
should 176
since 177
so 178
some 179
somehow 180
someone 181
something 182
sometime 183
sometimes 184
somewhere 185
still 186
such 187
than 188
that 189
the 190
their 191
them 192
themselves 193
then 194
thence 195
there 196
thereafter 197
thereby 198
therefore 199
therein 200
thereupon 201
these 202
they 203
third 204
this 205
those 206
though 207
through 208
throughout 209
thru 210
thus 211
to 212
together 213
too 214
top 215
toward 216
towards 217
under 218
until 219
up 220
upon 221
us 222
used 223
very 224
via 225
was 226
we 227
well 228
were 229
what 230
whatever 231
when 232
whence 233
whenever 234
where 235
whereafter 236
whereas 237
whereby 238
wherein 239
whereupon 240
wherever 241
whether 242
which 243
while 244
whither 245
who 246
whoever 247
whole 248
whom 249
whose 250
why 251
whyever 252
will 253
with 254
within 255
without 256
would 257
yes 258
yet 259
you 260
your 261
yours 262
yourself 263
yourselves 264

Appendix 4.6 1-word preambles

Actually	First	Only
Again	Furthermore	Oops
Ah	Gosh	Or
Alright	Here	Otherwise
Also	Hey	Quick
Also	However	Rather
And	Ie	Right
Anyhow	Incidentally	Second
Anyway	John	So
Besides	Lastly	Still
But	Like	Sure
Cripes	Look	Then
Damn	Maybe	Third
Definitely	Moreover	Though
Eg	Namely	What
Enough	Nevertheless	Whatever
Exactly	Next	Whenever
Excellent	No	Yeah
Except	Now	Yes
Fine	Okay	Yet

Appendix 4.7 2-word preambles

Enough gab	Just so
Er right	Last question
Oh dear	Moving on
Oh sure	Never mind
Oh right	No problem
Right fine	Not least
Sorry, Adam	Not really
Sounds alright	Now then
Very good	Of course
All right	Oh no
And er	Oh well
As well	Okay so
But then	Okay whatever
Even so	Right then
Fair enough	Really though
Fine but	So er
First off	So actually
For example	Something else
Go on	Tell me
Hold on	That said
I see	Then er
I mean	Well then
In addition	Yeah and
In brief	Yeah but
In fact	Yes definitely

Appendix 4.8 3-word preambles

Right, great, wonderful	Tell me this
While I think	Okay Adam so
To begin with	No not really
Oh goodness me	Right then so
Last of all	Oh yeah absolutely
First of all	Well perhaps but
I always wonder	Find out sometime
Kick this around	I see right
Right okay but	For Heaven's sake
Adam you first	Look here then
Just a second	Try this instead
Well I mean	To round off
But so anyway	Of course so
But I mean	Maybe somehow but
Wait a bit	The thing is
Changing the subject	For a start
I ask myself	All right so
Oh you know	Well maybe so
Oh I see	In your view
By the way	Yes, of course
But you know	In any case
One other thing	Just thinking aloud
Here hold on	Hm I wonder
Almost everyone asked	All the same
Hold your horses	Contribute for once

Appendix 5.1 Clustered Function Word Table

i 1	third 46	as 92
we 2	once 47	because 93
you 3	this 48	if 94
he 4	these 49	while 95
she 5	that 50	however 96
it 6	those 51	whenever 97
they 7	a 52	wherever 98
me 8	an 53	whether 99
us 9	the 54	whyever 100
him 10	all 55	thereby 101
her 11	alone 56	therein 102
them 12	another 57	thereupon 103
myself 13	any 58	thereafter 104
ourselves 14	both 59	whereafter 105
yourself 15	each 60	whereas 106
yourselves 16	either 61	whereby 107
herself 17	enough 62	wherein 108
himself 18	every 63	whereupon 109
itself 19	few 64	again 110
themselves 20	former 65	also 111
someone 21	latter 66	besides 112
anyone 22	last 67	moreover 113
noone 23	least 68	namely 114
everyone 24	less 69	hence 115
nobody 25	lot 70	so 116
something 26	lots 71	therefore 117
anything 27	many 72	thus 118
nothing 28	more 73	else 119
everything 29	most 74	instead 120
whoever 30	much 75	otherwise 121
whatever 31	neither 76	after 122
others 32	next 77	afterwards 123
mine 33	none 78	before 124
ours 34	only 79	meanwhile 125
yours 35	other 80	now 126
hers 36	several 81	then 127
my 37	same 82	until 128
our 38	some 83	anyhow 129
your 39	such 84	anyway 130
his 40	top 85	despite 131
its 41	whole 86	even 132
their 42	and 87	nevertheless 133
one 43	but 88	though 134
first 44	or 89	yet 135
second 45	nor 90	eg 136
	although 91	

ie 137	via 184	hereafter 230
per 138	with 185	hereby 231
re 139	within 186	herein 232
etc 140	without 187	hereupon 233
about 141	am 188	indeed 234
above 142	are 189	latterly 235
across 143	is 190	mostly 236
against 144	was 191	never 237
along 145	were 192	nowhere 238
among 146	be 193	often 239
amongst 147	been 194	out 240
amoungst 148	being 195	perhaps 241
around 149	became 196	somehow 242
at 150	have 197	sometime 243
behind 151	has 198	sometimes 244
below 152	had 199	somewhere 245
beside 153	do 200	still 246
between 154	does 201	there 247
beyond 155	did 202	together 248
by 156	done 203	well 249
down 157	will 204	almost 250
during 158	shall 205	rather 251
except 159	may 206	too 252
for 160	can 207	very 253
from 161	cannot 208	who 254
in 162	would 209	whom 255
inside 163	could 210	whose 256
into 164	should 211	what 257
near 165	might 212	which 258
of 166	ought 213	when 259
off 167	need 214	where 260
on 168	must 215	why 261
onto 169	used 216	how 262
outside 170	yes 217	whither 263
over 171	no 218	whence 264
since 172	not 219	
than 173	already 220	
thence 174	always 221	
to 175	anywhere 222	
toward 176	beforehand 223	
towards 177	elsewhere 224	
under 178	ever 225	
up 179	everywhere 226	
upon 180	formerly 227	
through 181	further 228	
thru 182	here 229	
throughout 183		

Appendix 5.2 Grouped Function Word Table

i 1	once 47	because 91
we 1	this 48	if 91
you 1	these 48	while 91
he 1	that 48	however 91
she 1	those 48	whenever 91
it 1	a 52	wherever 91
they 1	an 52	whether 91
me 1	the 52	whyever 91
us 1	all 55	thereby 91
him 1	alone 55	therein 91
her 1	another 55	thereupon 91
them 1	any 55	thereafter 91
myself 13	both 55	whereafter 91
ourselves 13	each 55	whereas 91
yourself 13	either 55	whereby 91
yourselves 13	enough 55	wherein 91
herself 13	every 55	whereupon 91
himself 13	few 55	again 110
itself 13	former 55	also 110
themselves 13	latter 55	besides 110
someone 21	last 55	moreover 110
anyone 21	least 55	namely 110
noone 21	less 55	hence 115
everyone 21	lot 55	so 115
nobody 21	lots 55	therefore 115
something 21	many 55	thus 115
anything 21	more 55	else 119
nothing 21	most 55	instead 119
everything 21	much 55	otherwise 119
whoever 21	neither 55	after 122
whatever 21	next 55	afterwards 122
others 21	none 55	before 122
mine 33	only 55	meanwhile 122
ours 33	other 55	now 122
yours 33	several 55	then 122
hers 33	same 55	until 122
my 37	some 55	anyhow 129
our 37	such 55	anyway 129
your 37	top 85	despite 129
his 37	whole 86	even 129
its 37	and 87	nevertheless 129
their 37	but 87	though 129
one 43	or 87	yet 129
first 44	nor 87	eg 136
second 45	although 91	ie 137
third 46	as 91	per 138

re	139		with	141		hereby	220
etc	140		within	141		herein	220
about	141		without		141	hereupon	220
above	141		am	188		indeed	220
across	141		are	188		latterly	220
against	141		is	188		mostly	220
along	141		was	188		never	220
among	141		were	188		nowhere	220
amongst		141	be	188		often	220
amoungst		141	been	188		out	220
around	141		being	188		perhaps	220
at	141		became		188	somehow	220
behind	141		have	188		sometime	220
below	141		has	188		sometimes	220
beside	141		had	188		somewhere	220
between		141	do	188		still	220
beyond	141		does	188		there	220
by	141		did	188		together	220
down	141		done	188		well	220
during	141		will	188		almost	220
except	141		shall	188		rather	220
for	141		may	188		too	220
from	141		can	188		very	220
in	141		cannot	188		who	254
inside	141		would	188		whom	254
into	141		could	188		whose	254
near	141		should	188		what	254
of	141		might	188		which	254
off	141		ought	188		when	254
on	141		need	188		where	254
onto	141		must	188		why	254
outside	141		used	188		how	254
over	141		yes	217		whither	254
since	141		no	217		whence	254
than	141		not	219			
thence	141		already	220			
to	141		always	220			
toward	141		anywhere		220		
towards		141	beforehand		220		
under	141		elsewhere		220		
up	141		ever	220			
upon	141		everywhere		220		
through		141	formerly		220		
thru	141		further	220			
throughout		141	here	220			
via	141		hereafter		220		