# The Development and Assessment of the Semantic Fields Model of Visual Salience.

Benjamin Stone, B.Soc.Sci. (Hons.)

School of Psychology, The University of Adelaide

*Thesis submitted for the degree of*

*Doctor of Philosophy*

February 2010

# Contents

Signed Statement . . . . . xviii

Acknowledgments . . . . . xx

Dedication . . . . . xxiii

Summary . . . . . xxiv

**Chapter 1. Introduction and Literature Review** **1**

1.0. Overview . . . . . 1

1.1. Display-based Web Page Search . . . . . 1

1.2. Semantic-based Web page Search . . . . . 6

1.3. Combining display-based and semantic-based information . . . . . 12

1.4. Introducing the Semantic Fields model . . . . . 12

1.5. Postlude . . . . . 16

**Chapter 2. Exegesis** **18**

2.0. Overview . . . . . 18

2.1. Paper 1: Validating pupil dilation as a measure of cognitive load . . . . . 18

2.2. Paper 2: Using LSA Semantic Fields to estimate visual salience on Web pages . 22

2.3. Paper 3: Improving the semantic component of the Semantic Fields model . . . 25

2.4. Paper 4: Assessing the improved Semantic Fields model estimates of visual salience on Web pages . . . . . 30

2.5. Summary . . . . . 36

2.6. Further notes on papers . . . . . 36

**Chapter 3. Pupil Size and Mental Load (2004)** **38**

**3.0. Abstract** **39**

**3.1. Introduction** **40**

3.1.1. The call for a measure of cognitive processing load . . . . . 40

3.1.2. Pupil dilation as a measure of cognitive load . . . . . 40

3.1.3. Pupil dilation as a measure of cognitive load in eye-tracking experiments that visually present stimuli on CRT monitors . . . . . 41

3.1.4. Experimental research hypotheses . . . . . 42

**3.2. Methodology** **43**

3.2.1. Participants . . . . . 43

3.2.2. Apparatus & Procedure . . . . . 44

**3.3. Results** **45**

3.3.1. Overall . . . . . 48

3.3.2. Rows . . . . . 48

3.3.3. Columns . . . . . 49

**3.4. Discussion** **51**

**Chapter 4. Using LSA Semantic Fields to Predict Eye Movement on Web Pages (2007)** **55**

**4.0. Abstract** **56**

**4.1. Introduction** **57**

4.1.1. Combining approaches . . . . . 57

4.1.2. Latent Semantic Analysis (LSA) . . . . . 58

4.1.3. LSA - Semantic Fields (LSA-SF) . . . . . 58

**4.2. Method** **59**

4.2.1. Participants . . . . . 59
4.2.2. Apparatus . . . . . 60
4.2.3. Procedure . . . . . 61
4.2.4. Calculating the LSA Semantic Fields . . . . . 63

**4.3. Results** **64**

4.3.1. Worst case to the best case scenarios . . . . . 65
4.3.2. Hyperlink-based LSA-SF . . . . . 66
4.3.3. All text LSA- SF . . . . . 67

**4.4. Discussion** **68**

4.4.1. Other sources of heat . . . . . 68
4.4.2. The addition of rules . . . . . 68
4.4.3. Benefits of the LSA-SF to eye-tracking research . . . . . 69
4.4.4. Summary . . . . . 70

**Chapter 5. Comparing Methods for Single Paragraph Similarity Analysis (in press)** **71**

**5.0. Abstract** **72**

**5.1. Introduction** **73**

5.1.1. Different types of textual language unit . . . . . 74
5.1.2. The dual focus of this paper . . . . . 78

**5.2. Semantic models, human datasets and domain-chosen corpora** **80**

5.2.1. Semantic models . . . . . 80
5.2.2. The Datasets . . . . . 83
5.2.3. Domain-chosen corpora: WENN (2000-2006) & Toronto Star (2005) . . . . . 85

**5.3. Study One. Comparison of models on domain-chosen corpora** **85**

5.3.1. WENN dataset & WENN Corpus . . . . . 86

5.3.2. Lee dataset & Toronto Star Corpus . . . . . 86

5.3.3. Summary of Study One . . . . . 89

**5.4. Study Two: Corpus Preprocessing** **90**

5.4.1. Removing numbers & single letters . . . . . 91

**5.5. Study Three: A better knowledge base?** **95**

5.5.1. Wikipedia Sub-corpora . . . . . 96

5.5.2. All models compared on Wikipedia sub-corpora . . . . . 101

**5.6. Study Four: Corpora that include the dataset paragraphs** **104**

**5.7. Overall Summary** **105**

**5.8. Discussion** **112**

**Chapter 6. Semantic Models and Corpora Choice when using Semantic Fields to Predict Eye Movement on Web pages (submitted)** **117**

**6.0. Abstract** **118**

**6.1. Introduction** **119**

6.1.1. Semantic Fields (SF) . . . . . 120

6.1.2. Focus of this paper . . . . . 120

**6.2. Method** **121**

6.2.1. Participants . . . . . 121

6.2.2. Apparatus . . . . . 122

6.2.3. Procedure . . . . . 123
6.2.4. Semantic Fields Models . . . . . 125
6.2.5. Corpora . . . . . 127
6.2.6. Baseline models to estimate eye-position . . . . . 129

**6.3. Results** **132**
6.3.1. Did the participants complete their tasks successfully? . . . . . 132
6.3.2. Were the participants paying attention? . . . . . 133
6.3.3. Ten models compared using the Bayesian Information Criterion . . . . . 136
6.3.4. How well does the VEC-SF model using the WIKI-WEB corpus predict the eye data? . . . . . 138

**6.4. Discussion** **140**

**6.5. Conclusions** **142**

**Chapter 7. General Conclusion** **143**
7.1. Final Statement . . . . . 148

**References** **149**

**A. Paper 1: Statement of contributions** **163**

**B. Paper 2: Statement of contributions** **167**

**C. Paper 3: Statement of contributions** **169**

**D. Paper 4: Statement of contributions** **172**

**E. Appendices from Paper 3 (Chapter 5)** **174**

E.1. Examples of similar and dissimilar paragraphs as rated by humans for the WENN dataset . . . . . 174

E.2. Examples of similar and dissimilar paragraphs as rated by humans for the Lee dataset . . . . . 176

E.3. Standard stop-list . . . . . 177

E.4. Corpora Parameters . . . . . 178

E.5. Study One results tables . . . . . 178

E.6. Stop-list used by Pincombe 2004 . . . . . 178

E.7. Study Two result tables . . . . . 180

E.8. IMDB-based Lucene query for Wikipedia . . . . . 181

E.9. Lee-based Lucene query for Wikipedia . . . . . 182

E.10. Study Three result tables . . . . . 183

E.11. Study Four results tables . . . . . 186

**F. Appendices from Paper 4 (Chapter 6) 188**

F.1. Goal pages with Semantic Field maps generated using Vectorspace and WIKI-WEB . . . . . 188

**G. Paper 1 - Original Article 197**

**H. Paper 2 - Original Article 204**

**I. Paper 3 - Original Article 211**

**J. Paper 3 - Supplementary Material file 272**

**K. Paper 4 - Original Article 301**

# List of Tables

2.1 Semantic Fields model with LSA and the TASA corpus, compared to Semantic Fields model with the semantic component held constant at one. The number of times higher Semantic Field values (all elements) were recorded for actual participant eye-points compared to eye-points generated in 1000 random trials. Best and Worst case calibration of eye-points are presented. . . . . 24

3.1 Descriptive statistics of the participants' Median Pupil Widths . . . . . 46

3.2 Approximate guide for converting Pixels x 10 into millimeters (mm). . . . . 47

3.3 Results of related samples t-tests used to compare average Median Pupil Widths recorded while participants were viewing the 'X' Stimulus in rows 0 to 4. 48

3.4 Results of related samples t-tests used to compare average Median Pupil Widths recorded while participants were viewing the 'X' Stimulus in columns 0 to 4. . . . . 49

4.1 Eye-based LSA-SF values compared to LSA-SF Overall Mean Values (OMVs) of the 1000 random trials associated with each page view. . . . . 65

6.1 Percentage of overlap between the expected landing Web page chosen by the experimenter and those chosen by the 49 participants. . . . . 133

6.2 Comparison of Bayesian Information Criteria (BIC) statistics calculated from log-likelihoods generated for all ten models. . . . . 137

E.4.1 Corpus parameters for the Toronto Star corpus, WENN corpus, and sub-corpora drawn from Wikipedia (1000 and 10000 documents) for both WENN and Lee datasets. . . . . 178

E.5.1 *t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, with the human ratings contained in the WENN dataset. None of the models' performance significantly improved when dimensionality was increased (alpha 0.05). Significant decreases in performance would be indicated by negative values equal to or greater than 1.96. So, in no case was increased dimensionality associated with significant decrements to model performance. . . . . 179

E.5.2 *t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, with human ratings contained in the Lee dataset. Significant decreases in performance would be indicated by negative values equal to or greater than 1.96. So, in no case was increased dimensionality associated with significant decrements to model performance. . . . . 179

E.7.1 Correlations (*r*) between similarity assessments of human raters and those made using LSA, Topic Model (Topics), Topic Model with Jensen-Shannon equation (Topics-JS), SpNMF at 50, 100, and 300 dimensions, and also the Overlap, Vectorspace and CSM models. The ALL columns display correlations based on corpora that contain both numbers and single letters (as used in Study One), conversely the NN-NSL columns are based on corpora with No Numbers and No single Letters (NN-NSL). Correlations exclude Same-Same document comparisons. . . . . 180

E.7.2 *t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, with human ratings contained in the WENN dataset used in Study Two. All corpora have had single letters and numbers removed. Significant decreases in performance would be indicated by negative values equal to or greater than 1.96. In no case was increased dimensionality associated with significant decrements to model performance. . . . . 181

E.7.3 *t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, with human ratings contained in the Lee dataset in Study Two. All corpora have had single letters and numbers removed. Significant decreases in performance would be indicated by negative values equal to or greater than 1.96. In no case was increased dimensionality associated with significant decrements to model performance. . . . . 181

E.10.1 Human to model correlations when estimating paragraph similarity on the WENN dataset, complex models using Wiki(pedia) 1000 & Wiki 10000 document corpora and the WENN Corpus (NN-NSL). Correlations exclude Same-Same paragraph comparisons. . . . . 184

E.10.2 Human to model correlations when estimating paragraph similarity on the Lee dataset, complex models using Wiki(pedia) 1000 & 10000 document corpora and the Toronto Star (NN-NSL) corpus. Correlations exclude Same-Same paragraph comparisons. . . . . 185

E.10.3 Examples of dimensions created by SpNMF on the 10000 document Wikipedia corpus generated for the Lee dataset where document length has been truncated at 100 words. . . . . 186

E.11.1 Comparison of models performance with standard Wikipedia 1000 corpora (Wiki 1000) and Wikipedia 1000 corpora including the 50 Lee paragraphs (Wiki 1050), using correlations between human and model estimates of paragraph similarity on the Lee dataset. Correlations exclude Same-Same paragraph comparisons. Significance tests were performed using Williams' T2 formula. . . . . 187

E.11.2 Comparison of models performance with standard Wikipedia 10000 corpora (Wiki 10000) and Wikipedia 10000 corpora including the 50 Lee paragraphs (Wiki 10050), using correlations between human and model estimates of paragraph similarity on the Lee dataset. Correlations exclude Same-Same paragraph comparisons. Significance tests were performed using Williams' T2 formula. . . . . 187

## List of Figures

1.1 Semantic Fields heat map of goal-oriented visual salience. Areas of greater estimated goal-oriented information salience have darker colors in this heat map. 13

2.1 Standardized pupil width during participants' fixations while they were performing 'add one' and 'substract seven' mathematical tasks. Time spent on each task (2 minutes) has been delineated into deciles. . . . . 21

2.2 Graphic representation of the display-based models and the Semantic Fields model used in Paper 4. Areas of greater estimated goal-oriented information salience have darker colors in these heat maps. . . . . 32

2.3 Standardized pupil width during participants' fixations while they were performing goal-oriented Web page navigation. Data is for all three Web sites. Time spent searching each page is delineated into deciles. . . . . 34

2.4 Semantic Field values (Vectorspace with the Wikipedia sub-corpus) calculated for participant eye-points during goal-oriented Web page navigation. Data is for all three Web sites. Time spent searching each page is delineated into deciles. 35

3.1 The layout of the experimental room. . . . . 43

3.2 The visual stimulus is randomly moved to another cell in the experimental display every five seconds. . . . . 44

3.3 A Boxplot illustration of the differences between participants' overall average SUB7 and ADD1 Median Pupil Widths. . . . . 47

3.4 A Boxplot illustration of the differences between participants' average SUB7 and ADD1 Median Pupil Widths in each of the 5 rows in the experimental grid. 49

3.5 A Boxplot illustration of the differences between participants' average SUB7 and ADD1 Median Pupil Widths in each of the 5 columns in the experimental grid. . . . . 50

3.6 A histogram displaying the relative position of the outliers in the distribution of ADD1 in C0. . . . . 51

4.1 Example of a LSA-SF Map with a participant's eye data super-imposed using black dots. . . . . 63

4.2 Eye LSA-SF average minus the LSA-SF OMV for each page viewed by participants using the Link-based LSA-SF method (Best Case). . . . . 66

4.3 Eye LSA-SF average minus the LSA-SF OMV for each page viewed by participants using the All Text-based LSA-SF method (Best Case). . . . . 67

5.1 Correlations ($r$) between the similarity ratings made on paragraphs in the WENN dataset by human raters and the those made by word overlap, LSA, Topics, Topics-JS (with Jensen-Shannon), SpNMF, Vectorspace, and CSM. All models, except word overlap used the WENN corpus. The effects of dimensionality reduction are displayed at 50, 100 and 300 dimensions for the more complex models that incorporate this reductive process. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons. . . . . 87

5.2 Correlations ($r$) between the similarity ratings made on paragraphs in the Lee dataset by human raters and the those made by word overlap, LSA, Topics, Topics-JS (with Jensen-Shannon), SpNMF, Vectorspace, and CSM. All models, except word overlap used the Toronto Star corpus. The effects of dimensionality reduction are displayed at 50, 100 and 300 dimensions for the more complex models that incorporate this reductive process. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons. . . . . 88

5.3 Correlations between similarity estimates made by human and models on paragraphs in the WENN dataset. Models that employ a knowledge base used the WENN corpus. “ALL” depicts standard corpus preprocessing used in Study One, “NN-NSL” corpora have also had numbers and single letters removed. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons. . . . . 93

5.4 Correlations between similarity estimates made by human and models on paragraphs in the Lee dataset. Models that employ a knowledge base used the Toronto Star corpus. “ALL” depicts standard corpus preprocessing used in Study One, “NN-NSL” corpora have also had numbers and single letters removed. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons. . . . . 94

5.5 Correlations between human judgments of paragraph similarity on the WENN dataset with estimates made using LSA (at 300 dimensions) using the WENN Wikipedia-based corpora containing 1000 and 10000 documents retrieved using Lucene with WENN-based query. Wikipedia documents have been truncated in four ways: first 100, 200, 300, and ALL words. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons. . . . . 99

5.6 Correlations between human judgments of paragraph similarity on the Lee dataset with estimates made using LSA (at 300 dimensions) using Lee Wikipedia-based corpora containing 1000 and 10000 documents retrieved using Lucene with Lee-based query. Wikipedia documents have been truncated in four ways: first 100, 200, 300, and ALL words. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons. . . . . 100

5.7 Correlations between human judgments of paragraph similarity on the WENN dataset with semantic model estimates made using Wikipedia Corpora with 1000 & 10000 documents and the WENN Corpus (NN-NSL). Error bars are the 95% confidence limits of the correlation. These results are also presented in Table E.10.1. Correlations exclude Same-Same paragraph comparisons. . . . 102

5.8 Correlations between human judgments of paragraph similarity on the Lee dataset with semantic model estimates made using Wikipedia Corpora with 1000 & 10000 documents and the Toronto Star (NN-NSL). Error bars are the 95% confidence limits of the correlation. These results are also presented in Table E.10.2. Correlations exclude Same-Same paragraph comparisons. . . . . 103

5.9 Correlations between human and model estimates of paragraph similarity on the Lee dataset using the standard Wikipedia 1000 corpora (Wikipedia 1000) and Wikipedia 1000 corpora including the 50 Lee documents (Wikipedia 1050). The overlap model has also been included in this bar graph to allow the reader another point of comparison. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons. . . 106

5.10 Correlations between human and model estimates of paragraph similarity on the Lee dataset using the standard Wikipedia 10000 corpora (Wikipedia 10000) and Wikipedia 10000 corpora including the 50 Lee documents (Wikipedia 10050). The overlap model has also been included in this bar graph to allow the reader another point of comparison. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons. . . . . 107

5.11 Scatterplots of the two best similarity estimates calculated for both the WENN and Lee datasets compared to the average similarity estimates made by humans for each pair of paragraphs. On the WENN dataset, (A) LSA using the WENN corpus (NN-NSL), and (B) the Overlap model. On the Lee dataset, (C) Vectorspace using the Wikipedia 1050 (including Lee documents), and (D) the Overlap model. Note, on the Lee dataset, average human ratings have been normalized [0,1]. . . . . 111

6.1 Semantic Fields Map using Vectorspace and a corpus drawn from Wikipedia. Participant's eye tracking data is super imposed using black dots. While the original SF model only used LSA, the SF models presented in this paper incorporate word overlap, Vectorspace, LSA, and SpNMF semantic models. Areas of greater estimated goal-oriented information salience have darker colors in this heat map. . . . . 121

6.2 Textual Web page elements are highlighted in red, images that have "ALT" or descriptive text are included. . . . . 131

6.3 Standardized pupil width during participants' fixations while they were performing goal-oriented Web page navigation. Time spent searching each page is delineated into deciles. . . . . 134

6.4 Semantic Field values (Vectorspace with the WEB-WIKI corpus) calculated for participant eye-points during goal-oriented Web page navigation. Time spent searching each page is delineated into deciles. . . . . 139

F.1.1 Mission Australia - Task 1, "Who is currently the Chief Operating Officer of Mission Australia?" Areas of greater estimated goal-oriented information salience have darker colors in this heat map. . . . . 188

F.1.2 Mission Australia - Task 2, "You are interested in working for Mission Australia. Search their Web site for the current job vacancies available at Mission Australia." Areas of greater estimated goal-oriented information salience have darker colors in this heat map. . . . . 189

F.1.3 Mission Australia - Task 3, "You are currently researching homelessness in young people and have heard that Mission Australia has recently published a report called 'The voices of homeless young Australians'. Search the Mission Australia Web site for this report into youth homelessness." Areas of greater estimated goal-oriented information salience have darker colors in this heat map. 190

F.1.4 Green Corps - Task 1, “You want to know more about Green Corps management. Find out who is the National Program Manager of Green Corps.” Areas of greater estimated goal-oriented information salience have darker colors in this heat map. . . . . 191
F.1.5 Green Corps - Task 2, “Find what environmental and heritage benefits are contributed by Green Corps.” Areas of greater estimated goal-oriented information salience have darker colors in this heat map. . . . . 192
F.1.6 Green Corps - Task 3, “Find the online Expression of Interest form to apply to become a Green Corps Partner Agency.” Areas of greater estimated goal-oriented information salience have darker colors in this heat map. . . . . 193
F.1.7 White Lion - Task 1, “Find out who is the current President of White Lion.” Areas of greater estimated goal-oriented information salience have darker colors in this heat map. . . . . 194
F.1.8 White Lion - Task 2, “You are interested in becoming a mentor for young people. Find out how to become one of White Lions mentors.” Areas of greater estimated goal-oriented information salience have darker colors in this heat map. 195
F.1.9 White Lion - Task 3, “You are interested in financial viability of White Lion as a business. Find out which Government Departments are supporters of the White Lion organization.” Areas of greater estimated goal-oriented information salience have darker colors in this heat map. . . . . 196

*Signed Statement*

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution to Benjamin Stone and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis (as listed below) resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the Web, via the University's digital research repository, the Library catalogue, the Australasian Digital Theses Program (ADTP) and also through Web search engines, unless permission as been granted by the University to restrict access for a period of time.

Stone, B. & Dennis, S. (submitted). Semantic models and corpora choice when using semantic fields to predict eye movement on web pages. *International Journal of Human-Computer Studies*.

Stone, B., Dennis, S., & Kwantes, P. J. (in press). Comparing methods for paragraph similarity analysis. *Topics in Cognitive Science.*

Stone, B., & Dennis, S. (2007). Using LSA semantic elds to predict eye movement on web pages. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th annual conference of the Cognitive Society* (pp. 665-670). Mahwah, NJ: Lawrence Erlbaum Associates.

Stone, B., Lee, M., Dennis, S., & Nettelbeck, T. (2004). Pupil size and mental load. *1st*

*Adelaide Mental Life Conference.* Available at:

http://www.psychology.adelaide.edu.au/cognition/aml/ Accessed April 2, 2009.

SIGNED: DATE:

*Acknowledgments*

Seven years is a long time to do anything in life, and certainly it is more than enough time to finish a Ph.D. thesis. It might go some way to helping the reader understand why this process has taken so long, if I mention that I have had as many supervisors as I have spent years completing this research.

In the beginning I worked with Dr Lynn Ward, Professor Ted Nettelbeck and Dr Brett Bryan. Both Lynn and Ted were from the School of Psychology and Brett was working with Geography department. Much of my first year was spent learning how to write computer programs. As an undergraduate, I had my first experience with computers' four years prior to this date, and it took until second year for me to worked up the confidence to submit an assignment that was not hand written. So, I had a fair way to catch up to improve my skill levels in this area. I thank Lynn, Ted and Brett for indulging me in this pursuit, as it has become a skill that I now use in my day to day work life and this thesis has relied heavily on these skills.

Brett is a good programmer and Geographical Information Systems expert. In my honors year we had developed a pupillometer using a video camera, and so the next logical step was to create the eye-tracker. Unfortunately, as was to become a recurring theme during my Ph.D., people have a life of their own to live, and the world does not revolve around me. At the end of my first year of candidature, Brett was offered a job at the CSIRO[1] and was unable to continue supervising me in this project. While I wish him all the best for this change in direction, it did stifle my plans to develop the eye-tracker and the Ph.D. thesis needed to be re-thought.

After taking a year off to work and plan a new Ph.D. project, I developed an interest in human behavior in Web based environments. Professor Michael Lee volunteered to supervise me in this new project, and Ted was kind enough to stay on as my secondary supervisor. The

[1] Australia's Commonwealth Scientific and Industrial Research Organisation.

approach of my research changed under Michael's supervision. Instead of developing an eye-tracker, Michael simply bought one. My research could now focus on the task of modeling users' behavior whilst engaged in Web tasks. After a year spent working with Michael, the University of California managed to enticed him away from Adelaide, and again I found myself without a supervisor. I also wish him all the best with his future pursuits.

Fortunately for me, Michael was replaced in the Psychology department by Dr Simon Dennis who was kind enough to take on the role as my principal supervisor. Simon really has been the driving force behind my academic development, and I will always be indebted to him for the friendship, guidance, and patience he has shown towards me. During the next two years in Adelaide, Simon helped develop my skills as both a programmer and a research scientist. However, as my story has already revealed, talented people are always in demand. Ohio State University offered Simon an Associate Professorship in 2007, and again I was without a primary supervisor in Adelaide. That said, Simon has continued to be my mentor in this research project as an external supervisor, and I thank him for sticking by me.

At the end of 2007, both Dr Dan Navarro and Professor John Dunn were kind enough to step into the roles of principal- and co-supervisors, respectively. I wish to thank them both for the support they have offered me over the last couple of years.

Many thanks go to Dr Peter Kwantes, who co-authored the third paper presented in this thesis with Simon and myself. Also many thanks go to the Defence Research & Development Canada (grant number – W7711-067985), who funded the research presented in the third paper. Furthermore, there have been numerous reviewers who have helped improve the four papers presented in this dissertation. I wish to extend my sincere appreciation for their helpful comments and suggestions, which have greatly improved the work present here.

So, seven years and as many supervisors on, I find myself at the end of this journey with a few others to thank. First, I would like to thank my wife Tegwen who is the most patient and understanding person I know. Next, my parents Barbara and Peter, who are both owed many

thanks for their constant encouragement and support. Finally, thank you to all my friends and colleagues who have had continued to encourage and motivate me towards the completion of this work.

*Dedication*

To those that I love most, Tegs, Mum, Dad, Chris and Del.

*Summary*

The present thesis describes the development and assessment of the Semantic Fields Model of visual salience. The Semantic Fields model provides estimates of visual salience in relation to goal-oriented Web site search tasks. The development and assessment of this model is reported over seven studies that are presented in two journal articles and two peer-reviewed conference papers.

In Paper 1 (N=50), pupil dilation is validated as a measure of cognitive load for use in later studies. While it has been found previously that a participant's pupil dilation will be larger during more complex tasks, these experiments have not generally been conducted under the environmental condition of light radiated from a computer monitor. The findings of this experiment indicate that computer monitor radiance in our experimental setting did not interfere with the ability to discriminate successfully between task-related pupil dilation.

Paper 2 (N=49) introduces the Semantic Fields model for estimating the visual salience of different areas displayed on a Web page. Latent Semantic Analysis and the Touchstone Applied Science Associates (TASA) corpus were used to calculate Semantic Field values for any (x, y) coordinate point on a Web page based on the structure of that Web page. These Semantic Field values were then used to estimate eye-tracking data that was collected from participants' goal-oriented search tasks on a total of 1842 Web pages. Semantic Field values were found to predict the participants' eye-tracking data.

In Paper 3 (N=100), four studies are present in which improvements are made to the semantic component of the Semantic Fields model. Estimates of textual similarity generated from six semantic models were compared to human ratings of paragraph similarity on two datasets. Results suggest that when single paragraphs are compared, simple non-reductive models (word overlap and vector space) can provide better similarity estimates than more complex models (Latent Semantic Analysis, Topic Model, Sparse Non-negative Matrix

Factorization, and the Constructed Semantics Model). Various methods of corpus creation were explored to facilitate the semantic models' similarity estimates. Removing numeric and single characters, and also truncating document length improved performance. Automated construction of smaller Wikipedia-based corpora proved to be very effective even improving upon the performance of corpora that had been chosen for the domain. Model performance was further improved by augmenting corpora with dataset stimulus paragraphs.

In Paper 4 (N=49), ten models are compared in their ability to predict eye-tracking data that was collected from participants' goal-oriented search tasks on a total of 1809 Web pages. Forming the basis of six of these models, three semantic models and two corpus types are compared as semantic components for the Semantic Fields model. Latent Semantic Analysis, Sparse Non-Negative Matrix Factorization, vector space, and word overlap were used to generate similarity comparisons of goal and Web page text in the semantic component of the Semantic Fields model. Vector space was consistently the best performing semantic model in this study. Two types of corpora or knowledge-bases were used to inform the semantic models, the well known TASA corpus and other corpora that were constructed from the Wikipedia encyclopedia. In all cases the Wikipedia corpora out performed the TASA corpora. The non-corpus based Semantic Fields model that incorporated word overlap performed more poorly at these tasks. Three display-based models were also included as a point of comparison to evaluate the effectiveness of the Semantic Fields models. In all cases the corpus-based Semantic Fields models outperformed the solely display-based models when predicting the participants' eye-tracking data. Both final destination pages and pupil data (dilation) indicated that participants' were actively performing goal-oriented search tasks.

Based on this research, it is concluded that the Semantic Fields model provided useful estimates of visual salience during participants' goal-oriented search of Web sites.

# Chapter 1. Introduction and Literature Review

## *1.0. Overview*

The research program presented in this thesis details the development and assessment of the Semantic Fields model of visual salience. Seven studies were performed that are reported in two journal articles and two peer-reviewed conference papers. These papers are presented in Chapters 3 to 6 in manuscript form, and their ordering reflects the chronology in which they were undertaken. The chapters surrounding these papers provide an introduction and literature review for the thesis (Chapter 1); an overview and explanation for the research that connects these papers in the context of the broader research program (Chapter 2); and the general discussion and conclusions of the thesis (Chapter 7). Some of the papers in this research required appendices, so these have been presented in the general appendices at the end of this manuscript.

Research focusing on visual salience and user behavior in Web page environments can generally be delineated into two main streams: display-based and semantics-based research. Display-based research focuses on the characteristics of elements displayed on the Web page (Grier, 2004; Ling & Van Schaik, 2004, 2002; Pan et al., 2004; Rigutti & Gerbino, 2004; McCarthy, Sasse, & Rigelsberger, 2003; Pearson & Van Schaik, 2003; Faraday, 2001, 2000). Alternately, semantics-based research has tended to focus on the semantic similarity between link text and search goal (Blackmon, Mandalia, Polson, & Kitajima, 2007; Fu & Pirolli, 2007; Blackmon, Kitajima, & Polson, 2005; Kaur & Hornof, 2005; Kitajima, Blackmon, & Polson, 2005, 2000; Cox & Young, 2004; Brumby & Howes, 2003; Chi et al., 2003; Pirolli & Fu, 2003; Blackmon, Polson, Kitajima, & Lewis, 2002).

## *1.1. Display-based Web Page Search*

### *1.1.1. Faraday's hierarchical cognitive model.*

Faraday (2000) purported a hierarchical cognitive model of user Web page search that combined both the characteristics of page elements and the area in which they are contained. He proposed that the inherent characteristics of the page elements govern their visual salience. More specifically, Faraday (2000) suggested that elements that move will be attended to before large elements, images, color, text style, and position, respectively. Faraday suggests that Gestalt principles govern the grouping of elements into separate areas of interest. For example, a collection of text hyperlinks could be grouped by either proximity or a similar background color. Upon initiating a goal-driven Web page search, the user is attracted to an element with the greatest visual salience (e.g., an animated Graphics Interchange Format image); she will then scan the area of interest that contains that element for target information. If the user is unsuccessful in satisfying their information need during this sequence, the element with the second greatest visual salience is located (e.g., a large heading) and its surrounding area of interest is scanned. This process is repeated until the user's goal is satisfied or the search is aborted.

In a series of eye-tracking experiments, Grier (2004) found little support for Faraday's model. While motion was found to attract initial user attention at levels greater than chance, there were no significant differences in participants' initial viewing preferences when size, image, color and text-style were manipulated. Grier criticized the lack of emphasis Faraday placed on position. Her research suggested that users predominantly focus initial attention to the middle of the screen, with the next most likely area of initial focus being the top-left part of the display (p.52). In Faraday's defense, after gaining more data from an eye-tracking study (Faraday, 2001), he revised some aspects of his model that were not examined in Grier's research. Furthermore, Faraday indicated that the content of the Web page would have "a strong impact upon the search and scan phases" (Section 8, 2000). However, while Faraday acknowledged its relevance, neither Faraday's nor Grier's work examined the affect that similarity between the semantic structures (or elements) of a Web page and the user's goals

produce in the Web page search process.

### *1.1.2. Position.*

Rigutti and Gerbino (2004) present the WebStep model that predicted differences in user behavior when following a navigation bar option as opposed to links embedded in the content text. They suggest users will judge that embedded links lead to specific information, whereas navigation links lead to broad categories. In their model, perceived distance to the target information moderates use of both the navigation (menu) bar and embedded links. When the distance was deemed to be small, their model predicts embedded links, as opposed to navigation bar links, are more likely to be chosen by the user. However, as the perceived distance to the target information increases, the probability that a navigation bar link will be chosen by user increases, and the probability that an embedded link will be chosen by the user decreases.

In the WebStep model, two other factors are predicted to influence the users decision to follow links. Firstly, utility of navigation bar links decrease as the position of the navigation bar is placed lower in the visual display. Secondly, headings increase the probability that embedded links will be selected by the user. However, the researchers do not specify how close the headings have to be either semantically or in proximity to the embedded link during this process. Rigutti and Gerbino do not deny the importance of semantic content in user link selection. However, they emphasized the importance of pragmatic or display-based models, such as WebStep, when semantic content alone does not discriminate user's choice in link following behavior.

McCarthy et al. (2003) utilized eye-tracking to explore optimal positioning of navigation menus. When comparing left, right and top positioned navigation menus, these researchers were surprised that most user gaze time was focused at the middle of the screen (48% of glances). More interestingly, they found that participants adapted to the Web page

environment. During second page views, users had increased (no inferential statistics were reported) the amount of gaze time on the region of the navigation menu regardless of its position (top, left or right of the page).

### *1.1.3. Learning*.

The notion of users adapting to, or learning, the structure of Web page environments is also supported by the research of Pan et al. (2004). In a study that compared user's eye movements over four types of Web sites (shopping, search, news, and business), results indicated a main effect for page order and an interaction effect between page order and Web site type on the dependent variables mean fixation duration, gazing time and saccade (occurrence) rate. The dependent variable 'mean fixation duration' was interpreted as a measure of task difficulty (Fitts, Jones, & Milton, 1950), while gazing time and saccade rate were both interpreted as being inversely related to task difficulty (Nakayama, Takahashi, & Shimizu, 2002).

Pan et al. had difficulties when interpreting their results, with apparent "contradictory conclusions" in the ocular measures that they used (p.5). It should be noted that Nakayama et al.'s tasks are quite different from those used in Pan et al.'s study. In Nakayama et al.'s study, saccade occurrence rate (the number of saccades per second) is measured while the participant visually follows an intermittently displayed target on a Cathode Ray Tube (CRT) monitor, and concurrently answers mathematical questions. In this situation, saccades could be prompted by search for the next target after the question has been answered. This is an occurrence that might happen less often in more difficult conditions that have longer calculation times, and thereby restrict the opportunity for users to perform after-task scanning. Moreover, this reduced opportunity to scan would also produce a negative relationship between task difficulty and saccade rate. Given the nature of Nakayama's task, it is not apparent that Saccade Rate can be interpreted in the same way when applied to the context of a user's Web page search.

In Pan et al.'s study, it is interesting to note that saccade (occurrence) rate only decreased from first page to second page viewed in the Search Web site condition. This could reflect the ordered structure of search engine output, which may reduce the need for user saccades when compared to the output of news, business, or shopping Web sites. At the very least, it signifies a need for researchers interested in the cognitive modeling of user search behavior to also consider the specific visual and semantic characteristics of the Web page, along with the user's environmental learning during navigation of a Web site.

### *1.1.4. Link color and Format.*

Ling and Van Schaik (2002) found that link color affected both visual search accuracy and speed in Web-page-like display. When searching for targets displayed in the default blue hyperlink on a white background participants results were less accurate than when the target link was presented in other formats such as: black link text on yellow backgrounds or black link text on white backgrounds. However, visual search times were generally faster when default colors were used. This may indicate that these participants were able to find and identify links more quickly when in default colors but, in a 'knee-jerk' like reaction, would choose to follow a link even if the content was wrong.

In a later study, Ling and Van Schaik (2004) manipulated font style of the hyperlinks presented in both navigation bars and the context areas of Web-page-like displays. Hyperlinks were presented in the default blue while other text was presented in black. Hyperlink font style was manipulated between normal, bold, underlined, and bold-underline conditions. These researchers report, that when hyperlink font style was manipulated in the navigation bar, there were no statistically significant differences in the participants' ability to correctly identify target hyperlinks. However, in the context area, the bold style was less accurately identified by participants when compared to the results of hyperlinks presented in either plain or bold-underlined styles. Ling and Van Schaik (2004) concluded that Web page designers need to

treat navigation bar and context area hyperlinks differently, ensuring that the latter are 'made more salient' to the Web page user (p. 919).

One problem with the experiments conducted by Ling and van Shaik is that the experimental stimulus lacked ecological validity. The Web-page-like displays were screen dumps (images) of Web pages that were displayed in Internet Explorer, but lacked basic functionality. That is, users could not mouse click on hyperlinks and traverse to new Web pages. Instead, responses were recorded with button presses, 'p' if the item was present and 'a' if the search item was absent from the display. Van Shaik notes the possibility that this display method changes the user's task, and that "automatic attention responses" Web users may display in their visual search behavior of Web pages may not be revealed under these simulated conditions (Pearson & Van Schaik, 2003, p. 345). Moreover, both authors propose that visual search processes of this type need to be evaluated "within the contexts of more realistic tasks" (Ling & Van Schaik, 2004).

#### *1.1.5. Section Summary.*

The display-based research reported above can be delineated into three broad Web page characteristics that influence a user's visual search of Web page hyperlinks. First, position has been shown to affect element salience and the likelihood of Web users' hyperlink following behavior. Second, aesthetic qualities of elements such as color and style, both attract the users attention and group hyperlinks into regions. Thirdly, environmental learning also appears to affect how Web users selectively focus attention during Web page navigation. Furthermore, eye-tracking has proved to be a useful tool in several research studies that investigate visual search for information in Web page environments.

### *1.2. Semantic-based Web page Search*

Semantic-based approaches to modeling user search behavior in Web pages emphasize the importance of hyperlink content and its similarity to user search goals. This similarity is

often referred to as the hyperlinks utility, 'information scent', or the perceived distance of the hypertext link from the users search goal. Two methods commonly employed for assigning a hyperlink's utility are: subjective assessment and automated semantic systems. Regardless of which of these methods is used to assign utility to the individual hyperlinks on a Web page, researchers contend that links with high utility will attract users' attention. Furthermore, if the link utility is above a set threshold (contingent to other predetermined rules), then the link clicking behavior will ensue.

Subjective assessment by human raters requires panels of judges to estimate the perceived degree of similarity between Web element text (e.g., a hyperlink) and a user's goal. The benefits of this method include a consistent degree of accuracy. For example, human raters can infer semantic similarity based on relatively little information provided by either the Web element's text or the user's goal. However, subjective assessments also have inherent problems. The time taken to make these ratings is both cumbersome and compromises any automation of Web page analysis. Furthermore, there may be idiosyncratic biases held by the raters (e.g., high IQ or greater technical vocabulary) which may not be also held by the everyday user.

Automated semantic systems can be divided into three categories: statistical, taxonomical, and hybrid (Kaur & Hornof, 2005). There are numerous examples of statistical systems: the vector space model (Salton, Wong, & Yang, 1975); Latent Semantic Analysis (LSA, Landauer, McNamara, Dennis, & Kintsch, 2007); the Topic model (Topics, Griffiths & Steyvers, 2002; Blei, Ng, & Jordan, 2003); Sparse Non-negative Matrix Factorization (SpNMF, Xu, Liu, & Gong, 2003); the Constructed Semantics Model (CSM, Kwantes, 2005); Hyperspace Analogue to Language (HAL, Burgess & Lund, 1997); Point-Wise Mutual Information using Information Retrieval (PMI-IR, Turney, 2001); and Non-Latent Similarity (NLS, Cai et al., 2004). Statistical systems or semantic models measure the probability of word co-occurrences from analysis of large collections of text documents. Alternatively, taxonomical systems, such as WordNet (G. A. Miller, 1995), create hierarchical databases in

which terms subsume related terms. For example, fruit would be a broader category containing oranges and apples. Hierarchical systems can also take into account word relations such as synonyms, hypernymy, and meronymy. Lastly, Hybrid systems such as *res* (Resnik, 1999) use both statistical and hierarchical methods to measure relation between terms.

*1.2.1. Subjective Assessment of Semantic Similarity and Web page navigation*. The Method for Evaluating Site Architectures (MESA) is a model of Web users' goal-oriented navigation and search times (Wu & Miller, 2007; C. S. Miller & Remington, 2004). Before MESA is used to predict Web site navigation, researchers first provide human assessments of utility for all Web page hyperlinks when compared to a given search goal. Using these human ratings of utility, MESA then conducts a serial assessment of the hyperlinks on the first or 'index' page of a Web site, using an 'opportunistic' strategy to follow the first hyperlink that exceeds a predetermined utility threshold. If no hyperlink on the index page exceeds this threshold, the threshold is then lowered and the hyperlinks are re-examined in the same manner. If still none of the hyperlinks meet this reduced criteria, then navigation is terminated. Alternatively, if a suitable hyperlink is found, it is followed and the process is repeated on the 'child' page. On the child page, if either the target information is not found, or there are no hyperlinks that exceed the threshold, then a 'back' operation is performed returning the model to the last 'parent' page.[1]

MESA is based on three principles: limited capacity, simplicity, and rationality. Like humans, MESA is 'limited' in the amount of information it can retain, and therefore only assesses one hyperlink at a time. MESA's evaluation of hyperlinks is 'simplistic', for example it evaluates all hyperlinks for the same amount of time regardless of their utility. The MESA model assumes the human users will exhibit 'rational' behaviour within the limitations of the first two principles outlined above. While MESA has successfully been used to model Web

[1]Note, a second assessment with lowered threshold is not performed on the hyperlinks for child pages.

users' navigation and search times, its lack of automated hyperlink utility assessment makes the MESA model difficult for researchers to implement on any scale. In more recent work, these reseachers have reported preliminary findings from a small (N=2) eye-tracking study that suggest easy search goals are associated with shorter fixation times. Wu and Miller (2007) propose that top-down knowledge of the search goal may be faciliating hyperlink processing and this is reflected by the participants' shorter fixations on these Web elements.

In an eye-tracking study, Brumby and Howes (2003) reported evidence that both past experience and utility of the surrounding menu options affected participants' goal-directed menu item search patterns. When goal items were situated amongst 'very bad' or lower relevance distracters, participants made fewer eye fixations on menu items than when 'moderate' or more relevant distracter items were used. Also, when an initial menu display was deemed to have been difficult, participants were observed to make more eye fixations in the next presented menu display, when compared to trials in which the initial menu had been deemed easy. Brumby and Howes (2003) also conclude that Web users' do not always access all of the menu options presented to them, and that users fixate on smaller subsets of menu items prior to selection of a item.

In a reanalysis of Brumby and Howes' (2003) data, Cox and Young (2004) propose a rational model of visual search through menu items. They suggest that menu item choice is dependent on the 'estimated relevance' or utility of previously viewed items in that menu (p.87). When surrounded by low utility distracter, high utility items that appear early in the menu's structure are likely to be immediately chosen, terminating the search process. However, if these items are presented towards the end of the menu structure, the model predicts visual search may not terminate with the high utility item on the 'first pass', but instead the menu items will be re-viewed, before possible termination of search and selection of the high utility item.

*1.2.2. Statistical Semantic Systems and Modeling of Web User Hyperlink Following Behavior.*

Grounded in the evolutionary-based theory of Information Foraging (Pirolli & Card, 1999), the Bloodhound Project models Web user behavior by accessing the 'proximal scent' (akin to 'information scent') of Web page links. Information scent is described as "the subjective sense of value and cost of accessing a page based on perceptual cues" (Chi, Pirolli, Chen, & Pitkow, 2001, p. 490). The Web User Flow by Information Scent (WUFIS) algorithm accesses un-normalized proximal scent of each hyperlink from proximal cues (e.g., hyperlink link text, surrounding text, page position), user information goals, and a Term Frequency by Inverse Document Frequency (TF.IDF) weighted matrix of all terms and documents contained in the Web site structure. These un-normalized values are then normalized so that all hyperlinks on a Web page sum to a value of one. When proximal scent cannot be calculated because the hyperlink is an image, distal scent is used instead. Distal scent is calculated in a similar fashion to proximal scent, except textual information from the 'linked to' page is combined with proximal cues in the algorithm. Chi et al. (2003) found that their modeling of Web user's link following behavior was very successful one-third of the time and moderately successful at other times.

Other research from the Palo Alto Research Center (PARC) has offered related models of users interaction with Web page environments. Pirolli and Fu's Scent-based Navigation and Information Foraging in the ACT architecture (SNIF-ACT) model combines Information Foraging Theory with the ACT-R model of cognition (Fu & Pirolli, 2007; Pirolli & Fu, 2003). Apart from its integration of ACT-R theory, SNIF-ACT also differs from its sibling, the Bloodhound project, in its use of a Bayesian-based model of 'spreading activation' (Pirolli, 1997) to access information scent (or link utility). SNIF-ACT's primary document corpus or knowledge base, the Tipster database, is augmented by the AltaVista search engine. AltaVista is used to expand the knowledge base of the SNIF-ACT model to incorporate 'popular' terms

that are not included in the Tipster database, such as movie titles and characters. Pirolli and Fu (2003) report that SNIF-ACT's modeling of user hyperlink clicking behavior exceeds what could be expected by chance. However, SNIF-ACT differs from the Bloodhound Project by primarily investigating and predicting when the user will leave a Web site, as information scent diminishes, in favor of pursuing other Web sites that may have more fruitful information sources.

The Cognitive Walkthrough for the Web (CWW) usability evaluation method is based on the Comprehension-based Linked model of Deliberate Search (CoLiDeS, Kitajima et al., 2005, 2000) for Web site navigation (Blackmon et al., 2007, 2005, 2002). Using CoLiDeS, CWW approaches the problem of modeling Web user's link following behavior in a somewhat similar fashion to the Bloodhound Project. Like the Bloodhound Project, some aspects of the screens display are taken into consideration, with screen areas being grouped into regions (navigation bar, tab menu, main content area, and so on). Also, like the Bloodhound Project, the semantic content of the each Web page is evaluated statistically against the Web user's target goals. However, instead of using the WUFIS algorithm, the CWW and CoLiDeS use LSA to compare semantic content. Furthermore, like SNIF-ACT, the CWW does not limit the content of its statistical semantic analysis to the documents in the Web site, but runs the LSA algorithm on a large corpus of documents that is considered to represent the user's knowledge base. Once a Web page has been segmented, the model generates a description of each section. These descriptions are then compared with the users goals using LSA and the knowledge base. The section description that has the highest similarity to these user goals will be selected for further analysis. Link texts in the selected section are then evaluated against the Web user's goal using LSA and the knowledge base. After this evaluation, the model then follows the hyperlink with the highest utility if its value is above a predetermined threshold.

### *1.2.3. Section Summary.*

Like the display-based studies reviewed above, semantic-based research has found both element position (within a menu) and environmental learning to be factors affecting Web user's hyperlink following behavior. Furthermore, users' Web navigation has been modeled with moderate success using automated statistical semantic-based systems that take some characteristics of Web page design into consideration.

### *1.3. Combining display-based and semantic-based information*

A review of both the display-based and semantics-based research into Web user's visual search and hyperlink clicking has indicated that the user's search processes are influenced by: text semantics, element position, aesthetic qualities of elements, and environmental learning. As is described above, Semantics-based researchers have, to varying degrees, started to incorporate characteristics of the Web page display into their models. Moreover, several researchers have highlighted the importance of this combined approach to modeling users navigation through Web sites (Blackmon et al., 2007, 2005, 2002; Fu & Pirolli, 2007; Kaur & Hornof, 2005; Chi et al., 2003; Pirolli & Fu, 2003).

### *1.4. Introducing the Semantic Fields model*

The Semantic Fields model attempts to estimate the visual salience of elements on a Web page relative to a Web user's information goal. It is based on the assumption that each element displayed on a Web page (such as hyperlinks, content text, and images) will influence how the Web page elements surrounding it are perceived and what value a Web page user will assign to them. The Semantic Fields model has two components: semantic and display. The semantic component provides estimates of similarity between a Web user's goal and the text[2] contained in each of the elements on a Web page. After the semantic component is calculated by the

[2]In its current form, the Semantics Fields model does not evaluate text contained within image files. Instead, in cases where a Web page designer has included descriptive text in the image tag's "alt" parameter, this semantic description of the image has been included in the Semantic Fields model's calculations. In future, Optical Character Recognition may be used by the Semantic Fields model to extract the text contained in image files.

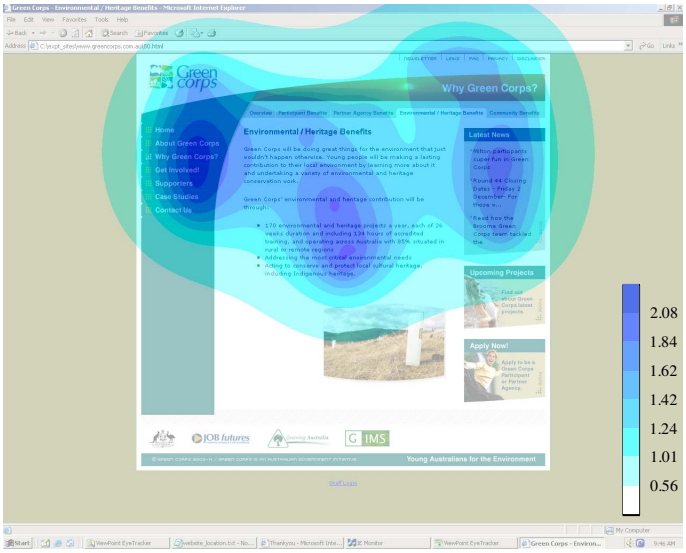



*Figure 1.1.* Semantic Fields heat map of goal-oriented visual salience. Areas of greater estimated goal-oriented information salience have darker colors in this heat map.

Semantic Fields model, the display component distributes and aggregates these estimates of semantic similarity over a Web page to provide estimates of the goal-specific visual salience. To illustrate, an example of the Semantic Fields estimates of goal-oriented visual salience of a Web Page is given in Figure 1.1.

### *1.4.1. Semantic Component.*

Similar to the CWW, LSA was initially chosen to inform the semantic component of the Semantic Fields model.[3] This decision was motivated by the successful implementation of LSA in tasks that require accurate estimates of the similarity between texts such as essay grading (Foltz, Laham, & Landauer, 1999). Furthermore, LSA has been shown to reflect human knowledge in a variety of ways. LSA measures correlate highly with humans' scores on standard vocabulary and subject matter tests; mimic human word sorting and category judgments; simulate word-word and passage-word lexical priming data; and accurately

[3]In the course of developing the Semantic Fields model other semantic models and corpora were also used to inform its semantic component (see Paper 4 in Chapter 6).

estimate passage coherence (Landauer et al., 2007).

In the first version of the Semantic Fields model, the well known Touchstone Applied Science Associates (TASA) corpus was selected to provide a generic knowledge base for LSA. The TASA corpus has been constructed to represent the reading material that is covered by American students up to their first year of college (Dennis, 2007). Therefore, it was thought that this knowledge base would be appropriate for use in modeling goal-oriented Web site searches performed by university staff and students in later experiments.

*1.4.2. Display Component: Combining the semantic component with knowledge of the display*.

One rationale underlying the concept of the Semantic Fields model, is that the information displayed on one region of a Web page, has an influence which goes beyond its immediate location. Moreover, this influence will decay as the point of focus is moved away from the information source. Using a decay function, LSA estimates of goal similarity are distributed and summed over each pixel position for all of the textual elements contained on a Web page (see Equation 1). Combining the semantic information ($L$) and its location with the distance ($d_{i(x,y)}$) from a given point $(x,y)$ using a decay function, enables the production of visual salience heat maps for Web pages (see Figure 1.1).

$$SF_(x,y) = \sum_i L_i e^{-\lambda d_{i(x,y)}} \quad (1)$$

A group of hyperlink elements that are close in spatial proximity may be recognized by the Web page user as a navigation menu. For instance, it is not that an individual hyperlink is recognized as a menu, but as Faraday (2000) suggests, it is the proximity and similarity in appearance to the set of hyperlinks around it, which forms their overall Gestalt or structure into a navigation menu. The Semantic Fields model captures this spatial relation between Web page elements. Such that, textual elements that are closely positioned on a Web page, such as

hyperlinks in a navigation menu, accumulate more utility or salience than items placed further apart.

Brumby and Howes (2003) report that Web users do not always access all of the menu options presented to them. Instead, the visual salience of a menu item is related to its goal-related utility, and the perceived utility of neighboring menu items. Similarly, the semantic component of the Semantic Fields model estimates the goal-related utility of individual menu items. The display component of the Semantic Fields model then distributes these estimates of goal-related utility among neighboring menu items.

Some researchers have suggested a general F-pattern is produced by Web users' reading patterns (Nielsen, 2006). However, it is unlikely that this type of model can accommodate navigation menus that are not located to the left of the display. Alternatively, the identification of menus by the Semantic Fields model is entirely governed by a Web page's element structure. Therefore, like the Web users' identified by McCarthy et al. (2003), the Semantic Fields model is able to accommodate varying menu locations.

As can be seen in Figure 1.1, varying degrees of visual salience have been estimated for content areas of the Web page. This varies because each element in the content (or non-menu) areas of the Web page will be assessed by the semantic component of the Semantic Fields model as holding varying similarity to the Web users' goal. As has been already suggested, this salience will also depend on the utility of the neighboring elements. Similar to the WebStep model proposed by Rigutti and Gerbino (2004), if goal-related embedded hyperlinks or other textual elements have neighboring elements that contain goal-related text (such as headings), then this will also increase the visual salience of that area.

### *1.4.3. Semantic Fields in comparison with the major models of Web site navigation.*

While Semantic Fields is not a model of Web site navigation that predicts a users hyperlink clicking, it is interesting to compare it to the major cognitive models of Web-

site navigation. Models like SNIF-ACT and MESA only assess the utility of individual hyperlinks against users' search goals when determining potential Web site navigation paths. Alternatively, models such as CoLiDeS and Semantic Fields both assess the collective utility of groups hyperlinks, and other Web page textual elements in the case of the Semantic Fields model. While Kitajima and colleagues have incorporated eye-tracking into their research (Habuchi, Kitajima, & Takeuchi, 2008; Habuchi, Takeuchi, & Kitajima, 2006; Namatame & Kitajima, 2008), they have not directly tested the hypothesis that the visual salience of Web page regions will be related to their semantic similarity to user goals. It is likely, that if CoLiDeS was to used to estimate visual salience on a Web page, Kitajama and colleagues would continue to segment a Web page into researcher defined regions. Alternatively, the generation of regions by the Semantic Fields model is completely automated, and built through knowledge of the relative positions of a Web page's textual elements. This feature of the Semantic Fields model makes it both simpler and more cost effective to implement than models like CoLiDeS.

### *1.5. Postlude*

Two areas of this introduction and literature review have been left for the reader to examine in the papers that make up the body of this thesis. All of the papers begin with a review of the relevant literature. In particular, literature on pupil dilation as a measure of cognitive load is left for Paper 1 (Chapter 3). Furthermore, literature on the semantic models which are used by the Semantic Fields model is more fully explicated in Paper 3 (Chapter 5). Motivating this decision to leave the introduction of these literatures for these papers is a desire to make this document less repetitive for the reader.

In the next chapter an Exegesis is presented that gives both an overview of each of the four papers presented in Chapters 3-6 and explains their individual contributions towards the development and assessment of the Semantic Fields model of visual salience. The exegesis

also contains findings discovered after the publication of some papers that have influenced the direction of this research project. Other information relating to the development of the Semantic Fields model that was considered important but outside the scope of the papers is also included in the Exegesis.

# Chapter 2. Exegesis

## *2.0. Overview*

The goal of this research project was to develop and assess the Semantic Fields model (see Section 1.4) of visual salience during goal-oriented Web site search tasks. Four papers are presented in Chapters 3-6 that document the progress of developing and assessing this model. In Paper 1 (see Chapter 3), pupil dilation is validated as a measure of cognitive load for use in later studies. Paper 2 (see Chapter 4) reports on the first attempt to use the Semantic Fields model to estimate data collected from 49 participants who were engaged in goal-oriented search on three Web sites. In Paper 3 (see Chapter 5), four studies are presented in which the semantic component of the Semantic Fields model was refined. Finally, in Paper 4 (see Chapter 6), the performance is compared between seven versions of the Semantic Fields model and three solely display-based models on the human goal-oriented Web site search dataset.

## *2.1. Paper 1: Validating pupil dilation as a measure of cognitive load*

This research program started with a validation of a measure cognitive load that was intended for use in later studies. For nearly 100 years researchers have reported that greater magnitudes of pupil dilation co-occurs when participants' are engaged in more difficult or cognitively demanding tasks. The human data in later research was to be primarily collected using eye-tracking software. So participants' pupil sizes recorded during experimentation could be measured concurrently with the (x,y) co-ordinates of the eye locations with little extra computational expense. Furthermore, in later studies the pupil dilation measure would be used to support the contention that participants were actively engaged in their Web site search tasks. Also, it was hoped that the pupil dilation measure might prove accurate enough to allow a micro analysis of participants' cognitive load during their Web page search tasks, and that this could then be compared to the estimates of information salience on each Web page provided by

the Semantic Fields model.

There are a large number of studies of cognitive abilities that have found that greater magnitudes of pupillary dilation are associated with more difficult tasks or greater cognitive load. Examples of these findings are found in literature on: short-term memory (Peavler, 1974; Kahneman & Beatty, 1966); long-term memory (Kahneman & Beatty, 1966); choice reaction time (Richer, Silverman, & Beatty, 1983); language processing (Hyönä, Tommola, & Alaja, 1995; Schluroff, 1982; Beatty & Wagoner, 1978); attention (Beatty, 1982); mathematical task complexity (Stone, Lee, Dennis, & Nettelbeck, 2004; Steinhauer, Condray, & Kasparek, 2000; Ahern & Beatty, 1979; Boersma, Wilton, Barham, & Muir, 1970; Schaeffer, Ferguson, Klein, & Rawson, 1968; Hess & Polt, 1964); and individual differences in cognitive ability (Ahern & Beatty, 1979; Peavler, 1974; Boersma et al., 1970).

Using the pupillary dilation as a measure of cognitive load is not without critics. Originally, the pupillary response was thought to be linked to an emotive response (Hess & Polt, 1960). This connection may have prompted other researchers to conclude that the increased magnitude of pupillary dilation during more difficult tasks was in fact a reflection of experimental arousal associated with undertaking more difficult tasks. It is therefore interesting to note, that Stanners, Coulter, Sweet, and Murphy (1979) who were initially proponents of an arousal explanation, concluded that:

> [t]he pupil response will show an arousal effect only when the cognitive demands of the situation are minimal. The control system is such that if the situation requires a substantial level of cognitive activity, only this will be indicated by the pupillary response. (p. 338)

In Paper 1 (see Chapter 3), a different problem is addressed that is associated with the implementation of the pupil measure; rather than its interpretation which is generally reported as a measure of cognitive or processing load. Authors such as Lin, Zhang, and Watson (2003)

failed to find an effect when comparing pupil dilations of participants' who were engaged in tasks of varying difficulty which were displayed on a computer monitor. Prior to using the pupil as a measure of cognitive load while participants were engaged in Web site search tasks, it was essential show that this effect could be found while participants' were using a computerized display.

To this end, an experiment was designed that required participants to complete mathematical tasks of varying levels of difficulty while viewing targets displayed on a computer monitor. The targets were presented in 25 evenly spaced cells of a five-five grid that covered all areas of the display. Participants were required to perform either add one or subtract seven calculations from a randomly generated number, while focusing on targets that were presented individually in the 25 cells. At one level, this experiment confirmed that task-evoked differences in pupillary size could be detected in all regions of the display. At another level, it enabled testing of factors that might have produced noise in the pupil data such as screen luminance and room lighting.

In the Paper 1, the difference between the medians of participants' pupil dilations during different tasks was used as the dependent variable. In hindsight, this was not the most appropriate measure to use. While the results are still valid, the participants' pupil data contained noise. A better comparison would have been made by standardizing each individual participants' pupil recordings. Calculating z-scores of participants' pupil sizes corrects for individual differences in pupil size and different levels of individual noise recorded for each participant. Figure 2.1 displays the average z-scores of the participants' pupil sizes during the two mathematical tasks, with the data delineated by time spent doing the task.

As can be seen in Figure 2.1, participants' pupil dilations are clearly larger during the difficult task. However, it is also interesting to note that over time participants' pupil average widths decrease in size. One possible interpretation is that as participants got better at the task, it became easier to do, and the decreased pupil dilation over time spent undertaking the task

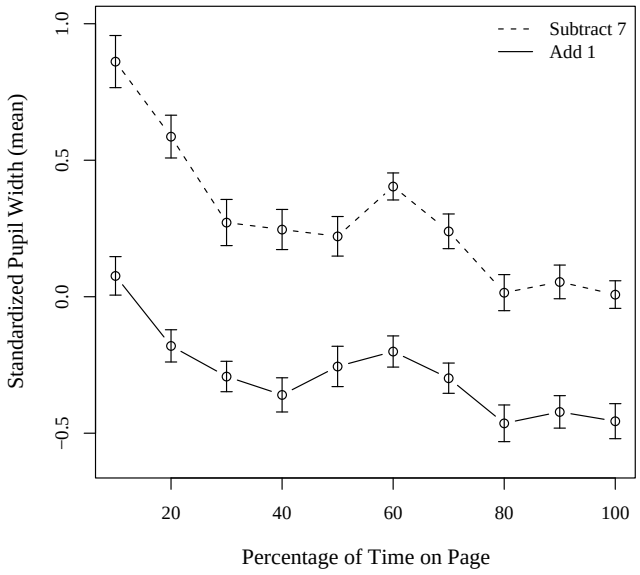



*Figure 2.1.* Standardized pupil width during participants' fixations while they were performing 'add one' and 'substract seven' mathematical tasks. Time spent on each task (2 minutes) has been delineated into deciles.

reflects a practice effect. However, completing a relatively simple task such as counting by ones would be unlikely to display large practice effects.

An alternative explanation is indicated by the rise in pupil size towards the middle of the time spent undertaking both tasks. It is likely that the overall decrease in pupil size is related to participants' waning performance over time. In this case, the smaller increase in pupil size during the middle of the task may reflect participants' briefly re-engaging with their task. Given the repetitive nature of the tasks, it is unsurprising that participants might become less motivated or lose track of their count as time spent undertaking these tasks increases. Furthermore, it was a desire to identify issues such as these, that motivated the researcher to validate the pupil as a measure cognitive load. With this measure in place, participants' pupil dilation were monitored in later experiments (see Chapter 6, Paper 4, Section 6.3.2) to check that participants were still "on task" during the goal-oriented Web page searches.

*2.2. Paper 2: Using LSA Semantic Fields to estimate visual salience on Web pages*

This paper describes the first attempt to use the Semantic Fields model to estimate the visual salience of different areas displayed on a Web page. To this end, 49 participants were asked to complete nine search tasks over three community services Web sites. In total, eye data were recorded during participants' goal-oriented search of 1842 Web pages.

Many eye-tracking studies of user behavior during Web-based tasks have lacked ecological validity (Ling & Van Schaik, 2004; Pearson & Van Schaik, 2003). For example, in some studies participants are asked to simulate Web browsing on non-interactive pages or screen dumps (Ling & Van Schaik, 2004). To address this problem, in this research participants were provided with a realistic setting (professionally developed Web sites) and familiar interface (Internet Explorer). To this end, a program was written by the candidate that would integrate Internet Explorer and the ViewPoint Eye-tracker for use in this research. Another issue, is that some experimental tasks do not extrapolate well to real world activities that are conducted by Web users (Ling & Van Schaik, 2004). Therefore, in this research, participants were given search tasks such as finding a board member of a company or a specific report on a Web site. Moreover, it was considered that these tasks could conceivably be performed during participants' personal research on the World Wide Web.[1]

The Semantic Fields model has been fully described in Section 1.4. of the introductory chapter. In brief, the model combines information about the semantic similarity of each Web page elements' text to a users' goals, with knowledge of those elements' position on that Web page. In Paper 2, the semantic component of the Semantic Fields model was implemented using Latent Semantic Analysis (LSA, Landauer et al., 2007) and the TASA corpus (see Section 4.1.2).

Two types of Semantic Fields models were examined. One model used the semantic

[1] A full list of these experimental tasks are presented in Section 6.2.3.

content from each Web page's hyperlinks to inform the semantic component of the Semantic Fields model. The other model used information collected from all of the textual elements on a Web page (including the text displayed in hyperlinks). The findings in this study indicated that both Semantic Fields models estimated participants' eye-movements better than a simulation which used randomized eye-points generated over 1000 trials. Moreover, the best results were obtained with the Semantic Fields model that used all of a Web page's textual elements, rather than only the text from hyperlinks on a Web page.

After Paper 2 (see Chapter 4) was published, an improved method of accessing where text is displayed on the Web page was implemented. The original method[2] of identifying the position of text was simply to use the (x,y) co-ordinates of an element's top-left corner, which could be retrieved from Internet Explorer Document Object Model. There are problems associated with using these points. Web developers have the option to align text in an element, however an element's size does not necessarily reflect the size of the text contained within it. For example, this problem can occur when a Web developer has created an element that spans the width of the Web page and contains text that is right aligned. In this case, the element (x,y) co-ordinates retrieved from Internet Explorer may be (0,0) or the top-left of the display, but the actual start location of the element's text may be (1100,0) which is in the top-right of the display.

To address the problem of identifying the true location of text on a Web page, a program was written to retrieve the dimensions of each textual element (x, y, width, height) from Internet Explorer. Using these dimensions, the program then extracted an image of each textual element from a screenshot of the Web page. Image processing was conducted on the element's image to identify the most common color (background), and the next most common color (text) within the image.[3] The center point (mid_x, mid_y) of the textual color was then used

[2]This method was used in Paper 2.

[3]This method may not be suited to all cases. It is conceivable that text could be displayed on a rainbow colored background. However, there were no colorful displays such as these in the Web sites used in this research.

to inform the Semantic Fields model of an accurate location of an element's text.

A further analysis of the data presented in Paper 2 revealed a problem with the semantic component of the Semantic Fields model. To test the semantic component, LSA values were held constant by setting them to one. If the semantic component was adding to the Semantic Fields model's performance, then it would be expected that the model would run more poorly in a condition which is only informed by the Web page element's position within the display. Surprisingly, including the semantic information provided only a marginal performance increase (2-3 percent)[4] to the Semantic Fields model (see Table 2.1).

Table 2.1: Semantic Fields model with LSA and the TASA corpus, compared to Semantic Fields model with the semantic component held constant at one. The number of times higher Semantic Field values (all elements) were recorded for actual participant eye-points compared to eye-points generated in 1000 random trials. Best and Worst case calibration of eye-points are presented.

| SF Model | Best case | % | Worst case | % |
|---|---|---|---|---|
| LSA & TASA corpus | 1817 | 98.64 | 1758 | 95.44 |
| No Semantic Model | 1775 | 96.36 | 1694 | 91.97 |

This analysis revealed that by using just the display component of the Semantic Fields model, it was possible to estimate participants' eye-movements with some degree of accuracy. That said, it is impossible for participants to find their search goals without reading the text displayed on the Web pages in this study. So, it was disappointing that the Semantic Fields model did not perform much better when the semantic information was included. One possibility was that LSA and the TASA corpus were not providing an adequate approximation of human knowledge required to assess similarity between the participants' goal and the textual information contained in the Web pages' elements. Therefore, to optimize the performance of the semantic component of the Semantic Fields model, four studies were

[4]It should be noted by the reader that the number of times the Semantic Fields model (using LSA and TASA) provided better estimates than the randomized trial model increased in comparison to the results presented in Paper 2. This increase in performance reflects the implementation of an improved method of accessing where text was displayed on the Web page.

conducted that examined various semantic models' ability to perform human-like similarity comparisons of text, and these studies are presented in Paper 3 (see Chapter 5).

Another possible problem with this assessment was that the randomized trials measure, that was compared to the Semantic Fields models, was too blunt an instrument to adequately delineate between model performance. This conclusion is supported by the high performance rates recorded for both the Semantic Fields models (both with semantic information included and without). To address this problem, the Semantic Fields models (with optimized semantic component) are more rigorously tested against solely display-based models in Paper 4 (see Chapter 6) using a Bayesian method of analysis.

### *2.3. Paper 3: Improving the semantic component of the Semantic Fields model*

The next goal in the research program was to improve the semantic component of the Semantic Fields model. There are several aspects of the semantic component that can be manipulated. For example, the semantic model can be changed, a different knowledge-base or corpus can be chosen, and the way in which the knowledge-base is pre-processed can be altered. In Paper 3 (see Chapter 5), these aspects of the semantic component were tested on two datasets of human similarity ratings for short paragraphs. Short paragraphs (50-200 words) represent the upper limit of the amount of text that would normally be displayed in a single element on a Web Page. Therefore, two datasets provided by Peter Kwantes and Michael Lee and colleagues, the WENN (see Section 5.2.2.1) and the Lee (see Section 5.2.2.2), seemed appropriate to use in this analysis.

#### *2.3.1. Study One: Comparing the performance of semantic models.*

In Study One, six semantic models were compared on their ability to estimate similarity between paragraphs. Their success in this task was measured by how closely these models' estimates correlated with similarity ratings made by human participants on WENN and Lee datasets. The models examined ranged in complexity. The simplest model, the word overlap

model, did not use a knowledge-base. Instead the word overlap model based estimates only on word co-occurrences in the paragraphs it compared. The other six models all used corpora as a knowledge-bases to inform their similarity estimates of paragraph similarity. These models were: vector space (Vectorspace, Salton et al., 1975); Latent Semantic Analysis (LSA, Landauer et al., 2007); the Topic model (Topics, Griffiths & Steyvers, 2002; Blei et al., 2003); Sparse Non-negative Matrix Factorization (SpNMF, Xu et al., 2003); and the Constructed Semantics Model (CSM, Kwantes, 2005). A description of each of these corpus-based semantic models is given in Section 5.2.1.

Two corpora were chosen to inform the corpus-based semantic models: the WENN corpus (for use on the WENN dataset) and the Toronto Star corpus (for use on the Lee dataset). Descriptions of these corpora are given in Section 5.2.3. The corpora contained short summaries or précises of media stories that were thought to represent the domain of knowledge humans may need to make similarity judgments on the paragraphs contained in the WENN and Lee datasets.

Surprisingly, in Study One, all of the corpus-based models performed very poorly when their estimates of paragraph similarity were compared to the human judgments contained in the WENN and LEE datasets. The only model to provide moderate correlations with the human data was the word overlap model (see Section 5.3). These results lead to three possible conclusions:

1. The models are unable to generate similarity calculations which are comparable with human judgments, and the Semantic Fields model would be better off using simple word overlap as its semantic component.

2. The pre-processing of corpora may have been inadequate, to the extent that noise had remained in the corpora which prevented the semantic models from making reasonable estimates of paragraph similarity.

3. Or, the corpora did not represent the knowledge required to make similarity estimates

on the paragraph contained in WENN and Lee document sets.

Given that models such as Vectorspace, LSA, Topics and SpNMF have been successfully implemented in a variety of tasks which require accurate estimates of the similarity between texts (Foltz et al., 1999; Griffiths & Steyvers, 2004; Berry & Browne, 2005; Salton et al., 1975), it is obviously unreasonable to accept that the models simply did not work. Therefore, the latter two possibilities are explored in Study Two and Study Three, respectively.

### *2.3.2. Study Two: Improving corpus pre-processing.*

Corpus pre-processing was manipulated in Study Two to examine the effects of removing numbers and single letters from the corpora used in Study One. This choice was inspired by examining the pre-processing techniques presented in Pincombe (2004) for the Lee dataset. Pincombe had found along with Lee and Welsh that LSA performed well at estimating paragraph similarity on the Lee dataset, with correlations between the model and human data (0.60) approaching the inter-rater reliability (0.605) reported in this study (Lee, Pincombe, & Welsh, 2005). So, it was interesting that their corpus pre-processing techniques precluded the inclusion of numbers and single letters. Possible justifications for the decision to remove these characters are discussed in the introduction to Section 5.4.

Including this method of corpus pre-processing improved nearly all of the models' performance on both datasets with the exception of CSM on the WENN dataset. In particular, model correlations with human similarity estimates on the WENN dataset were greatly improved. Both LSA (0.48) and SpNMF (0.43) performed well, providing moderate correlations to the human data. However, again to the surprise of the researchers, the performance of the word overlap had also increased with the removal of numbers and single letters, and still provide the best estimates of similarity on both datasets (WENN: 0.62, Lee: 0.53). Overall, the correlations were still low on the Lee dataset, which indicated that maybe the Toronto Star corpus was not proving to be a good match for the knowledge required to

estimate similarity on these paragraphs.

### *2.3.3. Study Three: Creating sub-corpora from Wikipedia.*

The poor performance of models using the Toronto Star corpus was problematic for the Semantic Fields model. This is because corpora that provide a good match to the knowledge required to make textual similarity judgments on Web sites are not always easy to obtain. Furthermore, this problem is only compounded by the great diversity of Web site content on the World Wide Web. Obviously, it is a difficult task to make a model that has the domain-specific background knowledge which enables it to successfully perform textual similarity comparisons on arbitrary Web sites. Therefore, a decision was made by the researcher to assess the performance of smaller domain-specific sub-corpora that were drawn from the larger collection of Wikipedia encyclopedia documents. The process by which these Wikipedia sub-corpora were created is outlined in Section 5.5.

Using 1000 document sub-corpora drawn from Wikipedia, a marked performance increase was observed for nearly all models on the Lee dataset. Furthermore, a corpus-based semantic model, Vectorspace (0.56) had finally performed better than the word overlap model (0.53) on the Lee-dataset. However, this increase in model performance was not found using the 1000 document Wikipedia sub-corpora on the WENN dataset when compared to the WENN corpus used in Study Two. In some ways this is not surprising, given that the paragraphs in WENN dataset are drawn from the larger collection of documents in the WENN corpus. As such, this larger collection of WENN documents provides a good approximation to the knowledge required to make similarity judgments in this case.

### *2.3.4. Study Four: Improving corpora by appending stimulus documents.*

In this final study, the stimulus paragraphs from the Lee dataset were appended to the Wikipedia sub-corpus that was generated for the Lee dataset. The rationale for undertaking this step is fully explained in Section 5.6. In brief, because the human raters had access to

all of the paragraphs contained in the datasets relatively early in their evaluation tasks, it seemed appropriate that the semantic models should be given a similar advantage. There are other reasons why this would be interesting for the Semantic Fields model. By adding the comparison documents into the corpus, it allows the semantic model to draw connections between words or phrases that may be unique to that stimulus paragraph. That is, a word or phrase may be contained in the paragraph dataset, however it may not be contained in the backgrounding corpora. The known words in the paragraph that co-occur with these unknown words allow the semantic model to place these unknown words in the correct context. In the case of the Semantic Fields model, if the textual information contained in each Web page is thought of as a paragraph or document, then these documents could be appended to the documents in a Wikipedia sub-corpus generated for a specific Web site. In the same way as is described for the Lee dataset, this would allow the semantic component of the Semantic Fields model to make associations between known and unknown words and phrases contained on a Web site. In the context of the Lee dataset, appending the experimental stimulus documents to the smaller 1000 Wikipedia sub-corpus again increased the performance of most models. The best performing models in this condition were now Vectorspace (0.67), LSA (0.60), SpNMF (0.56), and the word overlap model (0.53).

### *2.3.5. Paper 3 - Summary.*

In the context of the wider research framework, the studies presented in Paper 3 enabled the improvement of the semantic component of the Semantic Fields model. Prior to Paper 3, the Semantic Fields model used LSA and TASA to drive it's semantic component and make similarity comparisons between textual representations of Web search task and the text that was contained in Web Page elements. The findings in Paper 3 indicate that better estimates of textual similarity can be obtained using range of semantic models (not just LSA). These semantic models include word overlap, Vectorspace, LSA, and SpNMF. Also, it was shown

that when better corpora are not available, targeted Wikipedia sub-corpora can provide a good representation of the knowledge required to make textual similarity estimates. These corpora worked best when numbers and single letters are removed, and the sub-corpus size was restricted to 1000 documents. Performance was also improved by including the stimulus documents into the backgrounding sub-corpus. Based on these results, in the final paper (see Chapter 6), the semantic component of the Semantic Fields model used the best performing models in this study (word overlap, Vectorspace, LSA, and SpMF). Furthermore, these semantic models were used in conjunction with 1000 document sub-corpora drawn from the Wikipedia corpus, which had been pre-processed to remove numbers and single characters. Finally, text on each Web page viewed by participants was extracted as a document, and these documents were appended to the Wikipedia sub-corpora that were generated for each Web site.

### *2.4. Paper 4: Assessing the improved Semantic Fields model estimates of visual salience on Web pages*

Armed with improved methods of calculating the semantic component of the Semantic Fields model, it was now time to assess the improved model on the goal-oriented Web site search data used in Paper 2. Initial findings, outlined above, indicated that four main issues needed to be addressed in the current analysis. Firstly, it was desirable that the Semantic Fields model be compared against solely display-based models. This comparison would ensure that the semantic component of the Semantic Fields model was informing the model's estimates of the human eye-tracking data. Secondly, the randomized trials measure of performance was not able to adequately separate the performance of the Semantic Fields model and a display-based version of the Semantic Fields model which held the semantic component constant at one. Therefore, an alternative measure of performance was need in this final study. Thirdly, the semantic component of the Semantic Fields model could be improved, but this finding needed to be tested in the context of modeling the participants' eye-movements

during the goal-oriented Web site search tasks. Fourthly, it was apparent from the eye-tracking study performed in Paper 1, that participants may start their experimental tasks well, but that performance could wane over time. Therefore, the participants' ability to "stay on task" needs to be monitored. And, to this end, their pupil dilations would need to be monitored during experimentation.

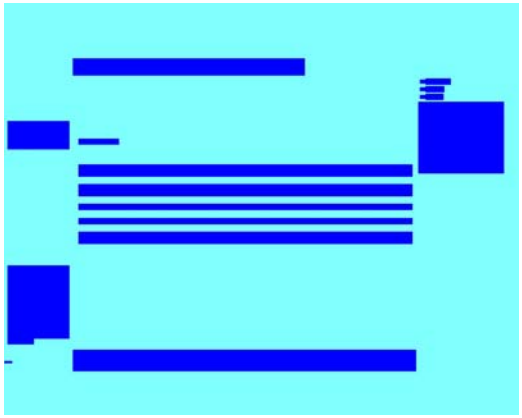
A new study was designed in which three display-based models were compared against seven Semantic Fields models on these models' ability to estimate the eye-movement data of 49 participants' during goal-oriented search tasks on a total of 1809 Web pages. Performance at this task was assessed by comparing the log-likelihoods of the ten models (see Equations 3-5 in Section 6.2.6) using the Bayesian Information Criterion (BIC, Schwarz, 1978) to compare these models' fit to the eye-tracking data.

The three display-based models are described in Section 6.2.6. To help the reader visualize the difference between these models and the Semantic Fields model, graphic representations of their heat maps (or information salience estimates) generated for one of the Web pages from the Mission Australia Web site (see Figure 2.2a) are displayed in Figure 2.2. The Flat model is the simplest of the display-based models and estimates that each point on the display has an equal probability of being viewed (see Figure2.2b). In the Non-Flat model, each textual element on the Web page is weighted as 3.41 times more likely to be viewed than non-textual elements (see Figure2.2c). This weighting reflects the optimized probability of an eye-point being in a text element, calculated over all Web pages viewed by participants for this sample (see Equation 4). The No-Model condition, is the Semantic Fields model with semantic values held constant at one (see Figure2.2d). Thus, the No-Model condition is informed by the relative position of the textual Web page elements, but not their semantic content. The Semantic Fields model has the same basic structure as the No-Model condition, however the semantic information given to the model has refined the distribution of heat within that basic structure (see Figure2.2e). Finally, in Figure 2.2f, the Semantic Fields model's estimation of

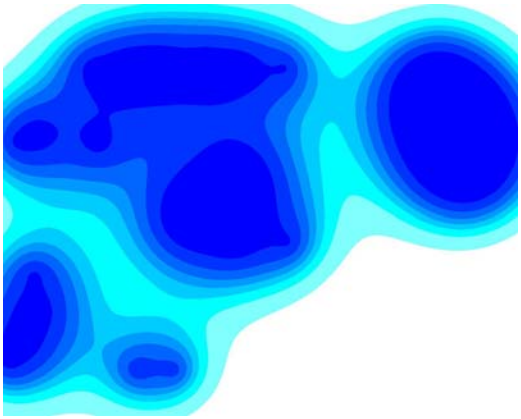
(a) A Mission Australia Web page.



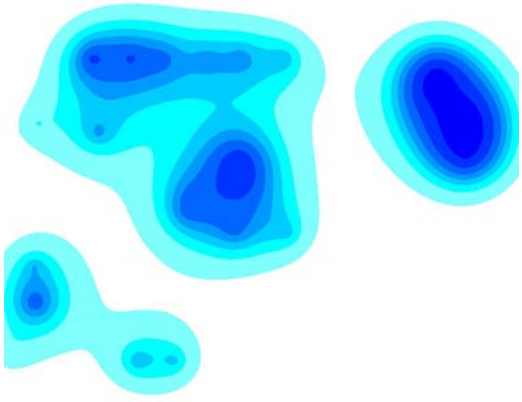
(b) Flat model.



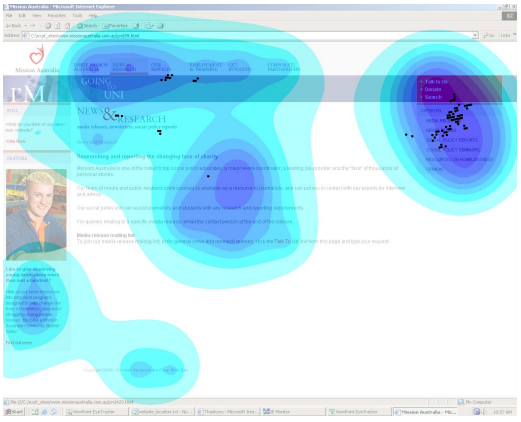
(c) Non-Flat model.



(d) No-Model (display-based - semantic component held constant at one in Semantic Fields model).



(e) Semantic Fields model.



(f) Semantic Fields model with a participants' eye-point superimposed in black.

*Figure 2.2.* Graphic representation of the display-based models and the Semantic Fields model used in Paper 4. Areas of greater estimated goal-oriented information salience have darker colors in these heat maps.

information salience is displayed with the Mission Australia Web page with a participants' eye movements superimposed onto the image.

Seven Semantic Fields models were used in this study to assess different configurations to the semantic component of this model. Six of these models used a corpus as a knowledge-base. These corpora were either the TASA or Web site specific sub-corpora that were drawn from Wikipedia. In line with the findings in Paper 3, the text contained on each Web page that was viewed by participants on a Web site was appended to the corpora used for that Web site as documents. Vectorspace, LSA and SpNMF were used in the semantic components of these corpus-based Semantic Fields models. This created a two by three design between corpora and semantic model. The seventh model used word overlap to inform the semantic component of the Semantic Fields model, and relied upon the co-occurrence of words between the search goal text and Web page element text.

A comparison using the Bayesian Information Criterion on log likelihoods calculated for each of the 10 models revealed that the Semantic Fields model with Vectorspace and the Wikipedia sub-corpora provided the best estimates of participants' eye movements during their search tasks. All of the corpus-based Semantic Fields models performed better than the display-based models. There were two main effects found in this analysis. Corpus choice appeared to affect the Semantic Fields model's performance. With the Semantic Fields models using the Wikipedia corpora outperforming the TASA corpora in all instances. The semantic models that were used in the Semantic Fields models also produced a main effect. Using either corpus as a knowledge base, Vectorspace was consistently the best performing semantic model, followed by SpNMF and then LSA.

Unlike the participants' undertaking the mathematical tasks in Paper 1 (see Figure 2.1), participants engaged in goal-oriented Web page search in this experiment appeared to be better motivated in their tasks. Pupil dilation data indicated that as the time participants spent searching a Web page for their goal increased, their cognitive load also increased (see

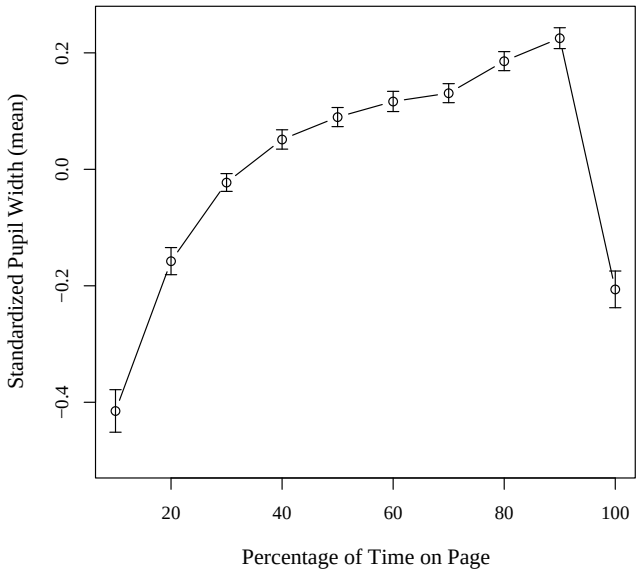



*Figure 2.3*. Standardized pupil width during participants' fixations while they were performing goal-oriented Web page navigation. Data is for all three Web sites. Time spent searching each page is delineated into deciles.

Figure 2.3). A likely explanation for the sharp drop in pupil size in the last 10 percent of time spent on a page, is that participants had either found their goal or had made the decision to click and move to another Web page. As noted in Paper 4, the conclusion that participants' were actively engaged in their tasks is also supported by the very high frequency with which participants found the target information set by the researcher on each Web site (see Section 6.3.1).

It had been found that the participants' were actively engaged in the search tasks on the three Web sites, and the Semantic Fields model that used Vectorspace and the Wikipedia sub-corpus performed best at estimating the participants' eye-points. But, how well was the Semantic Fields model doing at estimating the location of participants' focal points during these search tasks? It has been suggested that the display-based models such as the Non-Flat

and No-Model conditions appeared to offer a reasonable benchmark for comparison. However, to accomplish their tasks, participants engaged in visual search must 'seek out' locations of goal-relevant information on the display.

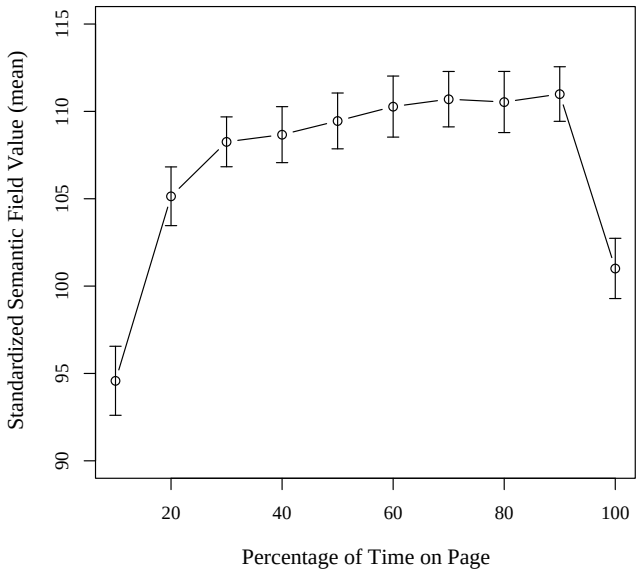



*Figure 2.4*. Semantic Field values (Vectorspace with the Wikipedia sub-corpus) calculated for participant eye-points during goal-oriented Web page navigation. Data is for all three Web sites. Time spent searching each page is delineated into deciles.

If the Semantic Fields model is able to capture some of the variability in participants' eye movements, then one would expect that Semantic Field values located at the participants' focal points will increase as participants' spend more time viewing a page. As can be seen in Figure 2.4, this was in fact the case. On all three Websites, it appears that after an initial orienting phase, the participants' eye-movements move towards areas of greater Semantic Fields values. As mentioned in the paragraph above, it is possible that the sharp drop off in the last 10 percent of time spent on the page captures eye movement around the time when participants' have click on a hyper-link and the browser is moving onto the next page.

### *2.5. Summary*

Both Web page semantics and display characteristics determine the success with which a user will be able to find information on a Web page. The Semantic Fields model incorporates both of these characteristics, and was found to provide better estimates of participants' eye-movements during goal-oriented search than could be generated by solely display-based models. Choices of both the semantic model and knowledge base affected the performance of the semantic component that is used by the Semantic Fields model. Contrary to expectations, a relatively simple semantic model, Vectorspace, outperformed more complex semantic models that employ dimensionality reduction. Also, better approximations to the knowledge required to successfully estimate textual similarity were produced by targeted sub-corpora drawn from Wikipedia when compared to those found using the more generic TASA corpus. Overall, the Semantic Fields model that used both Vectorspace and a targeted corpus drawn from Wikipedia, was found to be the best performing model when estimating participants' eye movements during goal-oriented search tasks in this research.

### *2.6. Further notes on papers*

It was not my original intention to undertake a Ph.D. by publication. However, on reflection it presented a good opportunity to publish my research work, rather than extracting and revising sections of a completed thesis at a later date. It has also provided me with excellent feedback by way of the peer-review process, and this has undoubtedly improved the overall caliber of the work presented here.

The work presented in these papers is predominantly my own. However, I should note here that the two datasets of human textual similarity judgments used in Paper 3 were not collected by me. The WENN dataset was collected by the Peter Kwantes from Defence Research and Development Canada. And the Lee Dataset was kindly provided by Michael Lee and colleagues. The source of these datasets is also acknowledged in Paper 3. Also, the

SEMMOD [5] package which was used to access the semantic models in all aspects of this research was created by my supervisor Simon Dennis and myself.

Different journals obviously have different requirements regarding the formatting of published work. Therefore, the formatting of the papers presented in the chapters of this thesis have been modified to conform to the overall style of this document. Section numbering has also been added to some papers to allow the reader easier access to sections of the papers. Where papers have included appendices or supplementary material, these sections have been added to the final appendices of this thesis. Copies of the original papers are also included in the appendices of this thesis (see Appendices G-K).

[5] The SEMMOD semantic models package was used to incorporate the Vectorspace model, Latent Semantic Analysis, and Sparse Nonnegative Matrix Factorization into the Semantic Fields model. The SEMMOD semantic models package is released under the GNU License and can be found here: http://mall.psy.ohio-state.edu/wiki/index.php/Semantic_Models_Package_%28SEMMOD%29

# Chapter 3. Pupil Size and Mental Load (2004)

Stone, B., Lee, M., Dennis, S., & Nettelbeck, T.

School of Psychology, University of Adelaide

1st Adelaide Mental Life Conference.

Statement of Contributions

Benjamin Stone (Candidate)

I was responsible for the conception and primary authorship of the paper. I was responsible for the development and programming of software based assessment tools and collection of all data. I conducted the statistical analyzes independently with advice from the co-authors. I was corresponding author and primarily responsible for responses to reviewers and revisions to the paper.

SIGNED: DATE: 10/8/10

Michael Lee (Co-author)

(see Appendix A)

Simon Dennis (Co-author)

(see Appendix A)

Ted Nettelbeck(Co-author)

(see Appendix A)

Stone, B., Lee, M., Dennis, S. & Nettelbeck, T. (2004). Pupil size and mental load. *1st Adelaide Mental Life Conference, Adelaide, S.A.*

# Chapter 4. Using LSA Semantic Fields to Predict Eye Movement on Web Pages (2007)

Benjamin Stone and Simon Dennis

School of Psychology, University of Adelaide

Proceedings of the 29th annual conference of the Cognitive Society (pp. 665-670)

Statement of Contributions

Benjamin Stone (Candidate)

I was responsible for the conception and primary authorship of the paper. I was responsible for the development and programming of software based assessment tools, and the collection and modeling of all data. I conducted the statistical analyzes independently with advice from the co-author. I was corresponding author and primarily responsible for responses to reviewers and revisions to the paper.

SIGNED: DATE: 10/8/10

Simon Dennis (Co-author)

(see Appendix B)

Stone, B. & Dennis, S. (2007). Using LSA Semantic Fields to Predict Eye Movement on Web Pages.
In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 665-670). Austin, TX: Cognitive Science Society.

# Chapter 5. Comparing Methods for Single Paragraph Similarity Analysis (in press)

Benjamin Stone

School of Psychology, The University of Adelaide

Simon Dennis

Department of Psychology, Ohio State University

Peter J. Kwantes

Defence Research and Development Canada (Toronto)

in press, Topics in Cognitive Science.

Statement of Contributions

Benjamin Stone (Candidate)

I was responsible for the conception and primary authorship of the paper. I was responsible for the development of the research program and modeling of all data. I conducted the statistical analyzes independently with advice from the co-authors. I was corresponding author and primarily responsible for responses to reviewers and revisions to the paper.

SIGNED: DATE: 10/8/10

Simon Dennis (Co-author)

(see Appendix C)

Peter J. Kwantes (Co-author)

(see Appendix C)

## 5.0. Abstract

The focus of this paper is two-fold. First, similarities generated from six semantic models were compared to human ratings of paragraph similarity on two datasets - 23 World Entertainment News Network paragraphs and 50 ABC newswire paragraphs. Contrary to findings on smaller textual units such as word associations (Grifths, Tenenbaum, & Steyvers, 2007), our results suggest that when single paragraphs are compared, simple non-reductive models (word overlap and vector space) can provide better similarity estimates than more complex models (LSA, Topic Model, SpNMF, and CSM). Second, various methods of corpus creation were explored to facilitate the semantic models' similarity estimates. Removing numeric and single characters, and also truncating document length improved performance. Automated construction of smaller Wikipedia-based corpora proved to be very effective even improving upon the performance of corpora that had been chosen for the domain. Model performance was further improved by augmenting corpora with dataset paragraphs.

## 5.1. Introduction

> The rate at which man [sic] has been storing up useful knowledge about himself and the universe has been spiralling upwards for 10,000 years.
>
> – (Toffler, 1973, p. 37)

Nearly four decades later, Toffler's remark is perhaps even more relevant in today's internet-driven world. 'Information overload' may be regarded as pervasive in many professions, and filtering strategies such as the summarization of text are commonplace. Government leaders and company executives make informed decisions based on briefs or short summaries of complex issues, provided by department managers who have in turn summarized longer reports written by their staff. In academia, the abstract is used to provide an overview of a paper's contents, so that time-pressed researchers can filter and absorb information related to their fields of study. In many areas it is important to be able to accurately judge the similarity between two or more paragraphs of information.

Sorting and extracting useful information from large collections of these types of summaries can prove both overwhelming and time consuming for humans. In an attempt to address this issue, semantic models have been successfully employed at these tasks. For example, Latent Semantic Analysis (LSA, Landauer, McNamara, Dennis, & Kintsch, 2007) has been used grade student essay scripts (Foltz et al., 1999). Similarly, the Topic Model has been used to extract scientific themes from abstracts contained in the Proceedings of the National Academy of Sciences (Griffiths & Steyvers, 2004). In a surveillance application, nonnegative matrix factorization has been applied to the large ENRON email dataset to extract topics or themes (Berry & Browne, 2005). Other models such as the vector space model (henceforth called 'Vectorspace'), were originally designed to index (or order by relevance to a topic) large sets of documents (Salton et al., 1975).

Semantic models have also been shown to reflect human knowledge in a variety of ways. LSA measures correlate highly with humans' scores on standard vocabulary and subject matter tests; mimic human word sorting and category judgments; simulate word-word and passage-word lexical priming data; and accurately estimate passage coherence (Landauer et al., 2007). The Topic Model has proven adept at predicting human data on tasks including: free association, vocabulary tests, lexical decision, sentence reading and free recall (Griffiths, Tenenbaum, & Steyvers, 2007). Other models have been developed to reflect specific psychological processes. For example, the Constructed Semantics Model (CSM) was developed as a global-matching model of semantic memory derived from the MINERVA 2 architecture of episodic memory (Kwantes, 2005).

### *5.1.1. Different types of textual language unit*

When making similarity comparisons on textual stimuli with semantic models, several researchers have highlighted the need to delineate textual stimuli into different language units (Kireyev, 2008; Foltz, 2007; McNamara, Cai, & Louwerse, 2007; Landauer & Dumais, 1997). Past research has modeled human comparisons of similarity on four types of textual language units: words, sentences, single paragraphs and chapters or whole documents (Foltz, 2007).

#### *5.1.1.1. Word comparisons.*

Griffiths et al. (2007) found that the Topic Model outperformed LSA on several tasks including word association and synonym identification. Griffiths and colleagues compared performance by the Topic Model and LSA on a word association task using norms collected by Nelson, McEvoy, and Schreiber (1998). The study used 4471 of these words that were also found in an abridged[1] 37,651 document (26,243 word, 4,235,314 token) version of the Touchstone Applied Science Associates (TASA) corpus. Moreover, the TASA corpus was

[1] A standard stop-list was applied, and only words appearing 10 times or more were included in the final corpus.

used as a knowledge base for both the Topic Model and LSA. Two measures were employed to assess the models' estimates of word association. The first measure assessed central tendency, focusing on the models' ability to rank word targets for each word cue. The other measure assessed the proficiency of each model's estimate of the most likely target response for each word cue. Griffiths and colleagues found that the Topic Model outperformed LSA on both of these performance measures. Furthermore, they reported that both models performed at levels better than chance and a simple word co-occurrence model.

In another study, Griffiths et al. (2007) compared the Topic model and LSA on a subset of the synonym section taken from the Test of English as a Foreign Language™(TOEFL®). The TOEFL was developed in 1963 by the National Council on the Testing of English as a Foreign Language, and is currently administered by the Educational Testing Service®.[2] The synonym portion of TOEFL offers four multiple choice options for each probe word, Griffiths and colleagues only included items in which all five words also appeared in the aforementioned abridged version of the TASA corpus. Similarity evaluations between the probes and possible synonyms, revealed that the Topics model (70.5%) answered more of the 44 questions correctly than LSA (63.6%). Furthermore, the Topic Model (0.46) predictions captured more of the variance found in the human responses than LSA (0.3).

The Topic model is a *generative model* that assesses the probability that words will be assigned to a number of topics. One of the key benefits of this generative process is that it allows words to be assigned to more than one topic, thus accommodating the ambiguity associated with homographs (Griffiths et al., 2007). For example, using the Topic Model the word 'mint' may appear in a topic that contains the words 'money' and 'coins', and in another topic containing the words 'herb' and 'plants'. Griffiths et al. (2007) argue that this attribute gives the Topic Model an advantage over models like LSA which represent meanings of words as individual points in undifferentiated Euclidean space (p. 219-220).

[2] http://www.ets.org/

*5.1.1.2. Sentence comparisons.*

McNamara et al. (2007) used several implementations of LSA to estimate the relatedness of sentences. The human judged similarity of these sentences decreased from paraphrases of target sentences, to sentences that were in the same passage as target sentences, to sentences that were selected from different passages to the target sentences. Likewise, comparing sentences using a standard implementation of LSA and the TASA corpus, these researchers found estimates of similarity were greatest for paraphrases, then same passage sentences, with different passage sentences judged least similar. When human estimates were correlated with the LSA estimates of sentences similarity, it was found that a version of LSA that emphasized frequent words in the LSA vectors best captured the human responses. Subsequently, using data collected in the McNamara et al. (2007) study, Kireyev (2008) found that LSA outperformed the Topic Model at this task.

*5.1.1.3. Single paragraph comparisons.*

Lee et al. (2005) examined similarity judgments made by Adelaide University students on 50 paragraphs that were collected from the Australian Broadcasting Corporation's news mail service. These paragraphs ranged from 56 to 126 words in length, with a median length of 78.5 words. Lee and colleagues compared several models' estimates of similarity to the aforementioned human ratings. These models included word-based, n-gram and several LSA models. Using a knowledge base of 364 documents also drawn from the ABC news mail service, LSA under a global entropy function[3] was the best performing model, producing similarity ratings that correlated about 0.60 with human judgments in this study. LSA's result in this study was also consistent with the inter-rater correlation (approximately 0.605) calculated by these researchers.

More recently, Gabrilovich and Markovitch (2007) produced a substantially higher

[3]Dividing by the entropy reduces the impact of high frequency words that appear in many documents in a corpus.

correlation with the human similarity judgments recorded for the Lee paragraphs (0.72) using the model they developed, Explicit Semantic Analysis (ESA). The ESA model uses Wikipedia as a knowledge base, treating Wikipedia documents as discrete human generated concepts that are ranked in relation to their similarity to a target text using a centroid-based classifier.

Kireyev (2008) used LSA and the Topic Model to estimate similarity of pairs of paragraphs taken from 3rd and 6th grade science textbooks. It was proposed that paragraphs that were adjacent, should be more similar than non-adjacent paragraphs. Difference scores were calculated between adjacent and non-adjacent paragraphs for both grade levels, with higher scores indicating better model performance. While it was not stated whether one model significantly outperformed the other at this task, on average LSA (0.75) scored higher on the 3rd Grade paragraphs than the Topic Model (0.49). However, there was little difference between the two models on the 6th Grade paragraphs (LSA 0.33, Topic Model 0.34).

*5.1.1.4. Chapters or whole document comparisons.*

Martin and Foltz (2004) compared whole transcripts of team discourse to predict team performance during simulated reconnaissance missions. Sixty-seven mission transcripts were used to create the researcher's corpus (UAV-Corpus). LSA was used to measure the similarity between transcripts of unknown missions to transcripts of missions where performance scores were known. To estimate the performance of a team based on their transcript using LSA, an average performance score was calculated from the 10 most similar transcripts found in the UAV-corpus. Performance scores estimated using LSA were found to correlate strongly (0.76) with the actual team performance scores.

Kireyev (2008) compared the similarity estimates of LSA and the Topic Model using 46 Wikipedia documents. These documents were drawn from six different categories: sports, animals, countries, sciences, religion, and disease. While both models correctly found more similarity between within-category documents than across-category documents, Kireyev

(2008) concluded that LSA performed this task consistently better than the Topic Model.

### *5.1.2. The dual focus of this paper*

This paper describes the outcome of a systematic comparison of single paragraph similarities generated by six statistical semantic models to similarities generated by human participants. Paragraph complexity and length can vary widely. Therefore, for the purposes of this research, we define a paragraph as a self-contained section of 'news' media (such as a précis), presented in approximately 50 to 200 words.

There are two main themes that are explored in this paper. At one level it is an evaluation of the semantic models, in which their performance at estimating the similarity of single paragraph documents is compared against human judgments. As outlined above, past research has indicated that performance of some models is clearly better depending on which type of textual units were used as stimuli. For example, the Topic Model was shown to perform better than LSA in word association research, where the textual unit was at the single word level. However, inherent difficulties such as homographs that affect models like LSA at the word unit level, may be less problematic for assessments made on larger textual units (sentences, paragraphs, and chapters or whole documents). These larger textual units contain concurrently presented words that may be less ambiguous and are thus able to compensate for a model's inability to accommodate homographic words (Landauer & Dumais, 1997; Choueka & Lusignan, 1985).

Research has indicated that LSA performs well at the paragraph level (Lee et al., 2005), but there are other models that may perform equally well if not better than LSA at this task. Therefore, in this research we compare six models' efficiency at the task of modeling human similarity judgments of single paragraph stimuli, the models examined were: word overlap, the Vectorspace model (Salton, Wong & Yang, 1975), Latent Semantic Analysis (LSA, Landauer, McNamara, Dennis, & Kintsch, 2007), the Topic Model (Griffiths & Steyvers, 2002, Blei,

Ng, & Jordan, 2003), Sparse Nonnegative Matrix Factorization (SpNMF, Xu, Liu, & Gong, 2003), and the Constructed Semantics Model (CSM, Kwantes, 2005). Our evaluation of these models is tempered by factors such a model compilation speed, consistency of performance in relation to human judgments of document similarity, and intrinsic benefits such as producing interpretable dimensions.

At another level this paper explores the characteristics of the corpora or knowledge bases utilized by these models that may improve models' performance when approximating human similarity judgments. With the exception of the word overlap model, a good background knowledge base is essential to the models' performance. Past research has identified various aspects of corpus construction that affect the performance of the Pointwise Mutual Information (PMI) co-occurrence model on word-based tasks such as the TOEFL synonym test (Bullinaria & Levy, 2006). These factors included: the size and shape of the context window, the number of vectors included in the word space, corpus size and corpus quality. To address this issue, we have evaluated aspects of corpus composition, preprocessing and document length in an attempt to produce suitable background corpora for the semantic models.

To this end, four studies are described in this paper that examine the semantic models' performance relative to human ratings of paragraph similarity. In the first study, semantic models use domain-chosen corpora to generate knowledge spaces on which they make evaluations of similarity for two datasets of paragraphs. Overall, the models performed poorly using these domain-chosen corpora when estimates were compared to those made by human assessors. In the second study, improvements in the models' performance were achieved by more thoroughly preprocessing the domain-chosen corpora to remove all instances of numeric and single alphabetical characters. In the third study, smaller targeted corpora (sub-corpora) constructed by querying a larger set of documents (Wikipedia[4]) were examined to assess whether they could produce sufficient performance to be generally useful (Zelikovitz & Kogan,

[4]http://en.wikipedia.org/

2006). In many applications the hand construction of corpora for a particular domain is not feasible, and so the ability to show a good match between human similarity evaluations and semantic models' evaluations of paragraph similarity using automated methods of corpus construction is a desirable outcome. Furthermore, document length of the essay-like Wikipedia articles was manipulated to produce better approximations of human judgment by the semantic models. Finally, in the fourth study, several of the models were found to produce better estimates of paragraph similarity when the dataset paragraphs were included in the models' backgrounding corpus.

## 5.2. Semantic models, human datasets and domain-chosen corpora

### *5.2.1. Semantic models*

The semantic models examined were word overlap, the Vectorspace model (Salton, Wong & Yang, 1975), Latent Semantic Analysis (LSA, Landauer, McNamara, Dennis, & Kintsch, 2007), the Topic Model (Griffiths & Steyvers, 2002, Blei, Ng, & Jordan, 2003), Sparse Nonnegative Matrix Factorization (Xu, Liu, & Gong, 2003) and the Constructed Semantics Model (Kwantes, 2005).

Word Overlap: Simple word overlap was used as a baseline in this research. It is the only model that does not use a corpus or knowledge base. Instead, it is a word co-occurrence model. Term frequencies are calculated for each paragraph in the dataset, and similarities are then measured as cosines (see Equation 1) of the resulting paragraph vectors.

$$\cos\theta = \frac{\mathbf{v_1} \cdot \mathbf{v_2}}{\|\mathbf{v_1}\| \, \|\mathbf{v_2}\|} \tag{1}$$

The Vectorspace model (Salton, Wong & Yang, 1975): The Vectorspace model assumes that terms can be represented by the set of documents in which they appear. Two terms will

be similar to the extent that their document sets overlap. To construct a representation of a document, the vectors corresponding to the unique terms are multiplied by the log of their frequency within the document, and divided by their entropy across documents, and then added. Using the log of the term frequency ensures that words that occur more often in the document have higher weight, but that document vectors are not dominated by words that appear very frequently. Dividing by the entropy or inverse document frequency (IDF) reduces the impact of high frequency words that appear in many documents in a corpus. Similarities are measured as the cosines between the resultant vectors for two documents.

Latent Semantic Analysis (LSA, Landauer, McNamara, Dennis, & Kintsch, 2007): LSA starts with the same representation as the Vectorspace model - a term by document matrix with log entropy weighting.[5] In order to reduce the contribution of noise to similarity ratings, however, the raw matrix is subjected to singular value decomposition (SVD). The SVD decomposes the original matrix into a term by factor matrix, a diagonal matrix of singular values, and a factor by document matrix. Typically, only a small number of factors (e.g., 300) are retained. To derive a vector representation of a novel document, term vectors are weighted, multiplied by the square root of the singular value vector and then added. As with the Vectorspace model, the cosine is used to determine similarity.

The Topic Model (Topics, Griffiths & Steyvers, 2002, Blei, Ng, & Jordan, 2003): The Topic Model is a Bayesian approach to document similarity that assumes a generative model in which a document is represented as a multinomial distribution of latent topics, and topics are represented as multinomial distributions of words. In both cases, Dirichlet priors are assumed. The parameters of these models can be inferred from a corpus using either Markov Chain Monte Carlo techniques (MCMC, Griffiths & Steyvers, 2002) or variational Expectation Maximization (Blei, Ng, & Jordan, 2003). We implemented the former. Ideally, document

[5]The reader is directed to Martin and Berry (2007) for an example of how to create a term by document matrix for both the Vectorspace model and LSA.

representations should then be calculated by running the MCMC sampler over a corpus augmented with information from the new document. To do this on a document by document basis is impractical. In the first instance, we choose to run the sampler over the corpus and then average the word distributions to calculate topic distributions for novel documents. Later in the paper, we investigate the impact of this decision by running the sampler over an augmented corpus containing all of the dataset paragraphs.

To calculate the similarity of the topic distributions representing documents, we employed both the Dot Product (see Equation 2) and Jensen-Shannon Divergence (JSD, see Equation 3). While the Dot Product was employed for convenience, the JSD is a symmetric form of the Kullback-Leibler Divergence (D) which derives from information theory and provides a well motivated way of comparing probability distributions.

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i \tag{2}$$

$$JSD(P \parallel Q) = \frac{1}{2} D(P \parallel M) + \frac{1}{2} D(Q \parallel M) \tag{3}$$

$$where \quad M = \frac{1}{2}(P+Q)$$

$$and \quad D(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Sparse Nonnegative Matrix Factorization (SpNMF, Xu, Liu, & Gong, 2003): Nonnegative Matrix Factorization is a technique similar to LSA, which in this context creates a matrix factorization of the weighted term by document matrix. This factorization involves just two matrices - a term by factor matrix and a factor by term matrix - and is constrained to contain only nonnegative values. While nonnegative matrix factorization has been shown to create meaningful word representations using small document sets, in order to make it possible to apply it to large collections we implemented the sparse tensor method

proposed by Shashua and Hazan (2005). As in LSA, log entropy weight term vectors were added to generate novel document vectors and the cosine was used as a measure of similarity.

The Constructed Semantics Model (CSM, Kwantes, 2005): The final model considered was the constructed semantics model (Kwantes, 2005). CSM was developed as a global-matching model of semantic memory derived from the MINERVA 2 architecture of episodic memory. Therefore, CSM is unique in that it was created primarily as a cognitive model to explain the emergence of semantics from experience. To this end, CSM uses a retrieval operation on the contexts in which words occur to generate semantic representations. It operates by taking the term by document matrix (using just log weighting) and multiplying it by its transpose. Consequently, terms do not have to appear together in order to be similar as is the case in the Vectorspace model. Again terms are added to create novel document vectors and the cosine is used as a measure of similarity.

### *5.2.2. The Datasets*

Two datasets of human ratings of paragraph similarity were used in this study. The first, which we will refer to as the WENN dataset was composed of similarity ratings generated by subjects comparing celebrity gossip paragraphs taken from the World Entertainment News Network. The second dataset, which we will refer to as the Lee dataset, was archival data collected by Lee et al. (2005).

#### *5.2.2.1. The WENN dataset.*

Students who were recruited by advertising the experiment on a local university campus, along with employees of Defence Research and Development Canada - Toronto (DRDC), provided paragraph similarity ratings from 17 participants to form the WENN dataset. Participants were paid CA$16.69 for taking part in the study. Twenty-three[6] single

[6]Participants actually compared 25 paragraphs, however a technical fault made the human comparisons of two paragraphs to the rest of the paragraphs in the set unusable.

paragraphs were compared by participants that were selected from the archives of World Entertainment News Network (WENN) made available through the Internet Movie Database[7] (see Appendix E.1). Paragraphs were not chosen randomly. First, each paragraph was chosen to be approximately 100 words long. The median number of words contained in paragraphs in the WENN dataset was 126, paragraph lengths ranging from 79 to 205 words. Paragraphs were also chosen in such a way to ensure that at least some of the paragraphs possessed topical overlap. For example, there was more than one paragraph about health issues, drug problems, stalkers, and divorce among those represented in the stimuli.

Participants were shown pairs of paragraphs, side by side, on a personal computer monitor. Pairs were presented one at a time. For each pair, participants were asked to rate, on a scale of 0 to 100, how similar they felt the paragraphs were to each other. Participants were not given any instructions as to the strategy they should use to make their judgments. Once a similarity judgment had been made, the next pair was presented. Each participant rated the similarity of every possible pairing of different paragraphs for a total of 253 judgments. Pearson correlations were calculated between participants' pairwise comparisons of the paragraphs in the WENN dataset, the average of these correlation coefficients (0.466) indicates that there was only moderate inter-rater reliability for the WENN dataset.

*5.2.2.2. The Lee dataset.*

Lee et al. (2005) recorded observations of paragraph similarity made by 83 Adelaide University students to form the Lee dataset. The dataset consists of ten independent ratings of the similarity of every pair of 50 paragraphs selected from the Australian Broadcasting Corporation's news mail service (see Appendix E.2), which provides text e-mails of headline stories. The 50 paragraphs in the Lee dataset range in length from 56 to 126 words, with a median of 78.5 words. Pairs of paragraphs were presented to participants on a computerized

[7] http://www.imdb.com

display. The paragraphs in the Lee dataset focused on Australian and international "current affairs", covering topics such as politics, business, and social issues. Human ratings were made on a 1 (least similar) to 5 (most similar) scale. As mentioned above, Lee et al. (2005) calculated an inter-rater reliability of 0.605.

### *5.2.3. Domain-chosen corpora: WENN (2000-2006) & Toronto Star (2005)*

Two corpora were chosen to act as knowledge bases for the semantic models to allow similarity estimates to be made on the paragraphs contained in the WENN and Lee datasets. The larger set of 12787 documents collected from WENN between April 2000 and January 2006 was considered a relevant backgrounding corpus for the 23 paragraphs contained in the WENN dataset, this larger set of documents is henceforth called the WENN corpus. It was not possible to resource the original set of 364 headlines and précis gathered by Lee et al. (2005) from the ABC online news mail service. Therefore, in an attempt to provide a news media-based corpus that was similar in style to the original corpus of ABC documents used by Lee and colleagues, articles from Canada's Toronto Star newspaper were used. Moreover, the Toronto Star corpus comprised of 55021 current affairs articles published during 2005.

Initially, both corpora were preprocessed using standard methods: characters converted to lower case, numbers were zeroed (i.e., 31 Jan 2007 became 00 jan 0000), punctuation and words from a standard stop-list (see Appendix E.3) were removed, and words that appear only once in a corpus were also removed. Descriptive statistics for both the WENN corpus and the Toronto Star corpus are displayed in Appendix E.4.

## 5.3. Study One. Comparison of models on domain-chosen corpora

Comparisons made between all semantic models and human evaluations of paragraph similarity for both datasets are presented in the following two subsections of this paper. For the

more complex models (LSA, Topics and SpNMF) one must select a number of dimensions in which to calculate similarities. Performance is likely to be influenced by this choice, therefore in each case comparisons were made using 50, 100 and 300 dimensional models.

### *5.3.1. WENN dataset & WENN Corpus*

Using the WENN corpus, correlations between similarity ratings made by humans and the models on paragraphs in the WENN dataset were low (see Figure 5.1) for all models except the simple word overlap (0.43). Of the other models, CSM (0.26) and LSA at 50 dimensions (0.21) performed best. Using the Jensen-Shannon metric improved the performance of the Topic Model in all cases when compared to the dot product measure of similarity. It could be argued that both Vectorspace ($r = 0.17$, $t_{(250)} = 1.61$, n.s.[8]) and LSA at 50 dimensions ($r = 0.21$, $t_{(250)} = 1.05$, n.s.) performed as well as the CSM on this document set. For LSA, the Topic Model and SpNMF, increasing the dimensionality or number of topics did not significantly increase or decrease model performance at this task (see Table E.5.1).

### *5.3.2. Lee dataset & Toronto Star Corpus*

Again except for the word overlap (0.48), the correlations between similarity ratings made by human participants and the models on the paragraphs in the Lee dataset were very low (see Figure 5.2). CSM and SpNMF (300 dimensions) were the next best performing models, correlating 0.15 and 0.14 with human judgments, respectively. Also, Vectorspace had higher correlations than both LSA and the Topic Model using the dot product similarity measure. In 9 out of 12 possible comparisons, increased dimensionality produced significantly better estimates of paragraph similarity by models when compared to human ratings (see Table E.5.2).

[8]Two-tailed significance tests (alpha = 0.05) between non-independent correlations were performed with Williams' formula (T2) that is recommend by Steiger (1980).

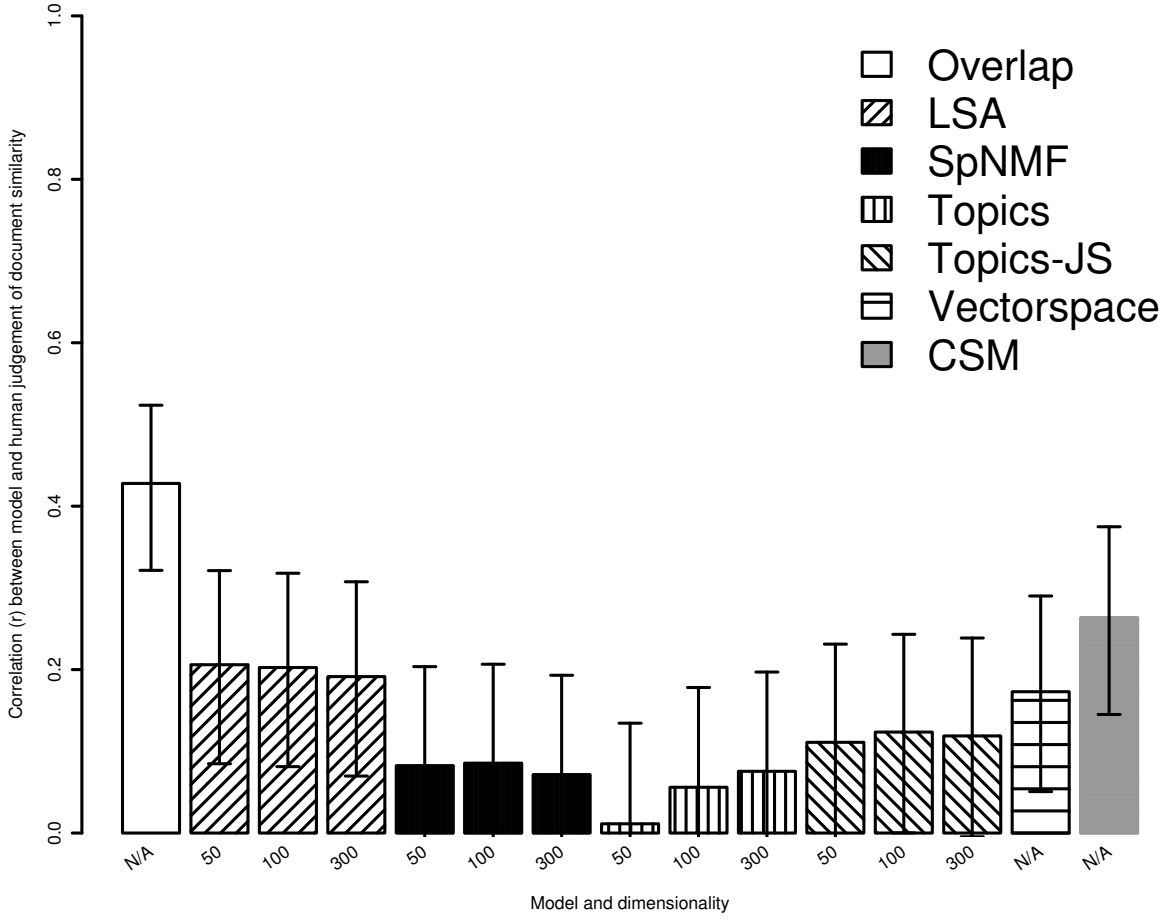



*Figure 5.1.* Correlations ($r$) between the similarity ratings made on paragraphs in the WENN dataset by human raters and the those made by word overlap, LSA, Topics, Topics-JS (with Jensen-Shannon), SpNMF, Vectorspace, and CSM. All models, except word overlap used the WENN corpus. The effects of dimensionality reduction are displayed at 50, 100 and 300 dimensions for the more complex models that incorporate this reductive process. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

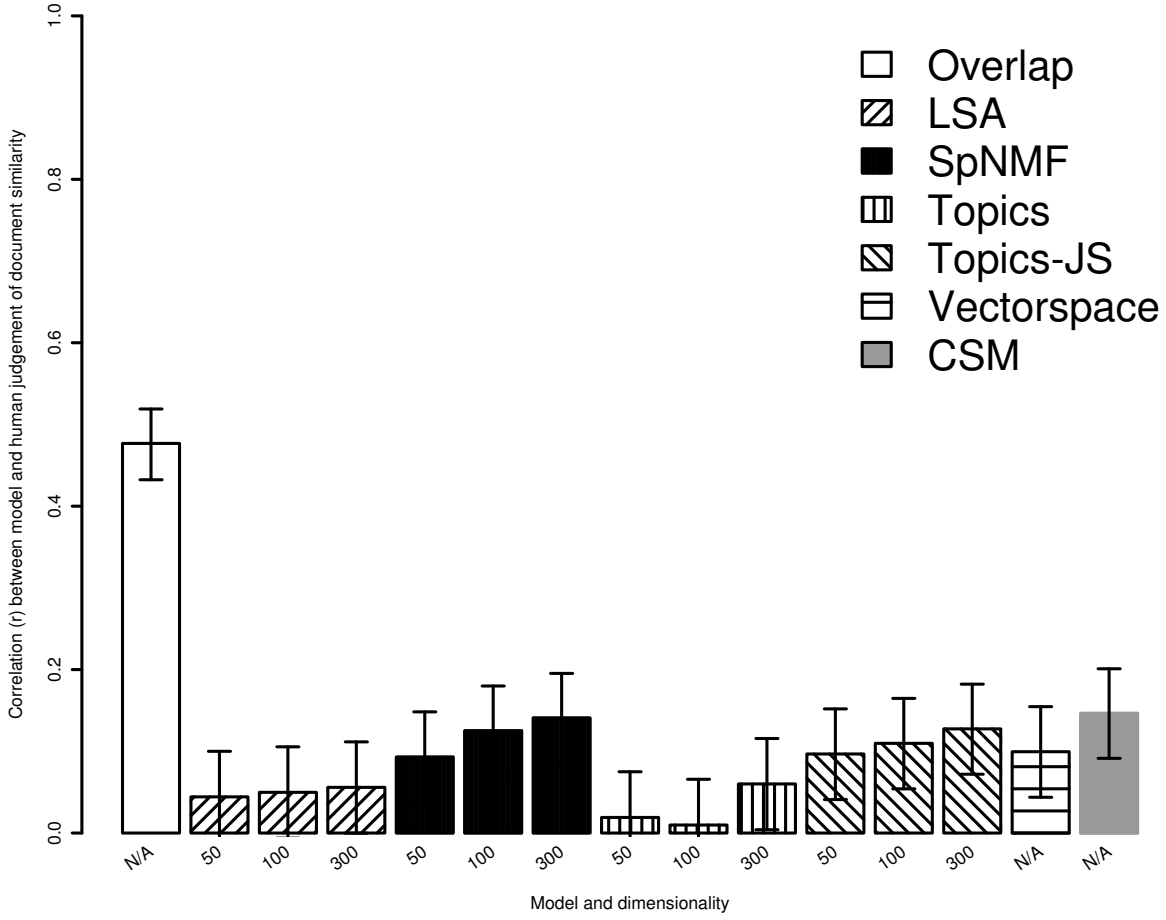



*Figure 5.2.* Correlations ($r$) between the similarity ratings made on paragraphs in the Lee dataset by human raters and the those made by word overlap, LSA, Topics, Topics-JS (with Jensen-Shannon), SpNMF, Vectorspace, and CSM. All models, except word overlap used the Toronto Star corpus. The effects of dimensionality reduction are displayed at 50, 100 and 300 dimensions for the more complex models that incorporate this reductive process. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

### *5.3.3. Summary of Study One*

Overall, the simple word overlap model outperformed the more complex semantic models when paragraph similarities were compared to human judgments made on both WENN and Lee datasets. On the Lee dataset, semantic models generally performed better when semantic spaces were compiled with higher dimensionality. However, when model dimensionality was increased on the WENN dataset, a similar increase in performance was not found. The generally poor results for the more complex models could be the product of at least one of the following circumstances:

a) the models are unable to generate similarity calculations which are comparable with human judgments.

b) the preprocessing of corpora may have been inadequate, to the extent that noise remained in the corpora which prevented the semantic models from making reasonable estimates of paragraph similarity.

c) or, the corpora did not represent the knowledge required to make similarity estimates on the paragraph contained in WENN and Lee document sets.

Other studies have reported more encouraging results when comparing estimates of paragraph similarity generated by semantic models and humans (Lee, Pincombe & Welsh, 2005 and Gabrilovich & Markovitch, 2007). Therefore, the first possible conclusion is likely to be inaccurate, indicating semantic models can make a reasonable estimate of the similarity of paragraphs when compared to human judgments. While this was not the case in this study, poor performance by the semantic models may have been driven by a sub-optimal match between the background corpus and the paragraphs being tested. The likelihood of this scenario is supported by the generally low correlations with human results obtained by all of the models that required a background corpus. The following three studies explore the latter two possibilities. In Study Two, a more stringent corpus preprocessing method is used to

improve on the results presented in Study One. In Study Three, Wikipedia is used to generate better backgrounding corpora, and this method again improves model estimates of paragraph similarity when compared to the human judgments. Then, in Study Four, paragraphs from the datasets are added to the models' knowledge base to again improve model performance at this task.

## 5.4. Study Two: Corpus Preprocessing

Generally, corpus preprocessing identifies words that are likely to be informative to the semantic model. In the field of information retrieval there have been many types of sophisticated term selection functions employed by researchers (Sebastiani, 2002, p. 15). Other methods such as employing a stop-list are less complex, requiring no mathematical calculation, and simply remove words from the corpus which are deemed uninformative by the researcher. Stop-lists are usually applied to remove words such as articles, pronouns and conjunctions (Moed, Glänzel, & Schmoch, 2004). Bullinaria and Levy (2006) found that stop-lists reduced model performance when the textual unit under comparison is at a word-word level (such as the TOEFL task described above). However, working with paragraph comparisons, Pincombe (2004) states that "[u]se of a stop word list almost always improved performance" when comparing models estimates of similarity and human judgments (p. 1). A closer inspection of the stop-list (Appendix E.6) and preprocessing techniques (p. 14) used by Pincombe (2004)[9] was conducted. This review revealed that single letters had been removed by the author and only alphabetical characters had been used in his corpora. The difference between the preprocessing used in Study One (allowing the inclusion of zeroed numbers and single characters) and that used in Pincombe's research begs the question:

Can the removal of single letters and numbers from the background corpus improve a semantic model's ability to estimate paragraph similarity?

[9]These techniques were also used in the Lee et al. (2005) study.

It is possible that the presence of these types of information (numbers and single letters) in a corpus can create noise for the models. For example, the American Declaration of Independence in 1776 has little to do with Elvis Presley's birthday in 1935. Although using the preprocessing method of zeroing numbers, models comparing texts that describe these two occasions would erroneously find some similarity between them. Moreover, the zeroing of the aforementioned dates could also suggest commonality with a document describing the distance between two cities, obviously creating noise in the corpus even if this new document described a 1000 mile drive between Philadelphia (Pennsylvania) and Tupelo (Mississippi). Similarly, the 'Js' in 'J F K' and 'J K Rowling' should not indicate semantic similarity between documents that make reference to these well known individuals. Therefore, the removal of these items may benefit a model's ability to perform similarity ratings between paragraphs.

### *5.4.1. Removing numbers & single letters*

All numbers and single letters were removed from both the WENN and Toronto Star corpora[10] to test the hypothesis that removing these characters would improve the semantic models' performance when similarity ratings were compared to human judgments. Figure 5.3 and Figure 5.4 display comparisons between the results generated in Study One (ALL) and the results for spaces compiled on corpora without number and single letters (NN-NSL, No numbers - No Single Letters). Only the results for models compiled at 300 dimensions (where dimensionality is a parameter of the model) are displayed in these figures. It should be noted, while the models compiled at 300 dimensions generally[11] produced the best results, models compiled at both 50 and 100 dimensions displayed an identical trend (see Table E.7.1) of better

[10] Both corpora had already been preprocessed with standard methods: removing stop-words, punctuation, and words that appear in only one document were also removed.

[11] With the exception of Topic Model using the Jensen-Shannon metric, all models that incorporate dimensionality reduction performed better at 300 dimensions. Topics-JS at 100 topics was 0.29 compared to 0.28 with 300 topics.

performance when using the more stringent preprocessing method.

Although it may seem counter intuitive to remove information from a knowledge base or corpus, the removal of numbers and single letters improved correlations between human judgments and similarity ratings produced from models in nearly all comparisons that were made for both the WENN (see Figure 5.3) and Lee (see Figure 5.4) datasets. The only model that did not improve in performance was CSM on the WENN dataset. This difference for CSM between ALL (0.26) and NN-NSL (0.16) corpora was significant ($t_{(250)} = -2.48$, $p < 0.05$). A more promising trend was displayed by the other models, especially on the WENN dataset with the LSA (0.48) and SpNMF (0.43) models performing best of the more complex semantic models. However, this trend was also displayed by the simple word overlap model which continued to clearly outperform the other models. When numbers and single letters were removed from the paragraphs used by the overlap model, correlations between this model and the human judgments improved to 0.62 on the WENN dataset and 0.53 on the Lee dataset. In 4 out of 12 comparisons on the WENN dataset, and 5 out of 12 comparisons on the Lee dataset, increased dimensionality led to significant improvements to models' performance (see Tables E.7.2 and E.7.3).

Notwithstanding this general improvement in the more complex semantic models' performance, correlations with human judgments of similarity were still low using the Toronto Star (NN-NSL) corpus on the Lee dataset, with the highest being the Vectorspace model (0.2). This suggests that while corpus preprocessing was hindering the models' ability to provide reasonable estimates of paragraph similarity, there are also other factors that are impeding the models' performance. Clearly, the information and themes contained within corpora certainly constrain the performance of semantic models. However, suitable knowledge bases are not always easy to obtain. In an attempt to address this issue, the third study examines an alternative method of generating corpora that draws sets of knowledge-domain related documents (sub-corpora) from the online encyclopedia Wikipedia.

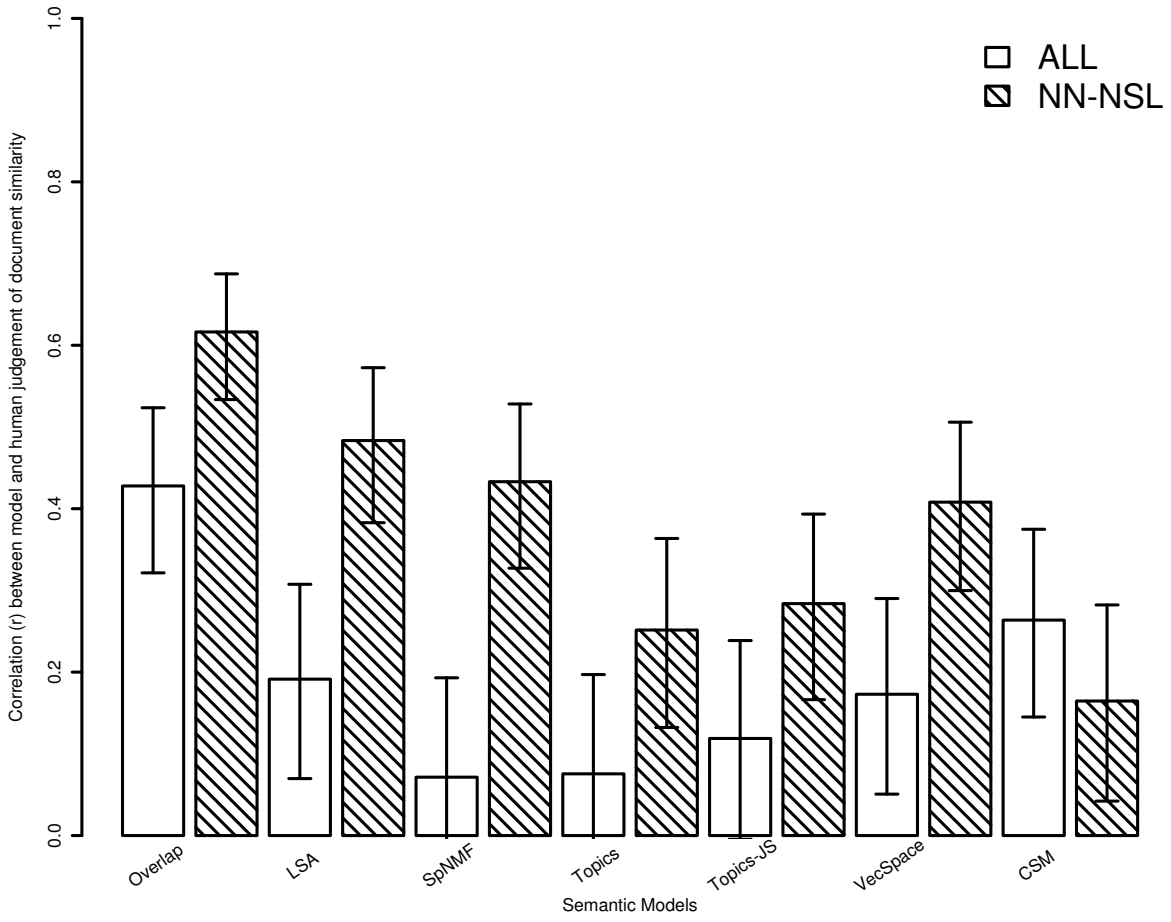



*Figure 5.3*. Correlations between similarity estimates made by human and models on paragraphs in the WENN dataset. Models that employ a knowledge base used the WENN corpus. "ALL" depicts standard corpus preprocessing used in Study One, "NN-NSL" corpora have also had numbers and single letters removed. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

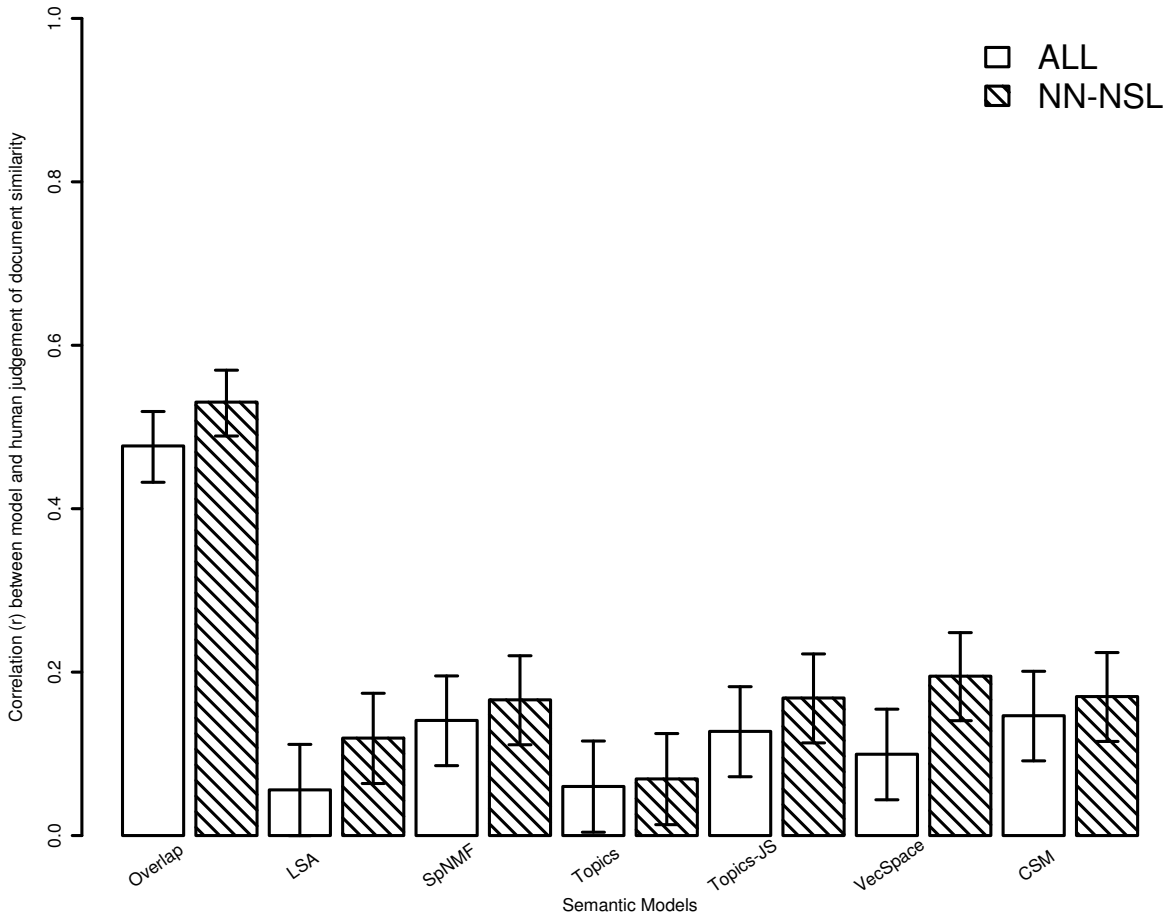



*Figure 5.4*. Correlations between similarity estimates made by human and models on paragraphs in the Lee dataset. Models that employ a knowledge base used the Toronto Star corpus. "ALL" depicts standard corpus preprocessing used in Study One, "NN-NSL" corpora have also had numbers and single letters removed. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

## 5.5. Study Three: A better knowledge base?

Smaller, more topic focused, sub-corpora may provide context for polysemous words, that may otherwise take on several meanings in a larger corpus. To this end, Wikipedia was utilized as a generic set of documents from which smaller targeted sub-corpora could be sampled and compiled. Wikipedia is maintained by the general public, and has become the largest and most frequently revised or updated encyclopedia in the world. Critics have questioned the accuracy of the articles contained in Wikipedia, but research conducted by Giles (2005) did not find significant differences in the accuracy of science-based articles contained in Wikipedia when they were compared to similar articles contained in the Encyclopedia Britannica. Furthermore, the entire collection of Wikipedia articles are available to the general public and can be freely downloaded.[12] All Wikipedia entries current to March 2007 were downloaded for this research. In total there were 2.8 million Wikipedia entries collected; however, the total number of documents was reduced to 1.57 million after the removal of incomplete articles contained in the original corpus. Moreover, incomplete articles were identified and removed if they contained phrases like “help wikipedia expanding” or “incomplete stub”. The resulting Wikipedia corpus was further preprocessed in the same manner as the NN-NSL corpora in Study Two: removing stop-words, punctuation, words that only appeared once in the corpus, and finally removing all numbers and single letters.

To enable the creation of sub-corpora, Lucene[13] (a high performance text search engine) was used to index each document in the Wikipedia corpus. Lucene allows the user to retrieve documents based on customized queries. Like the search results provided by Google, the documents returned by Lucene are ordered by relevance to a query.

Targeted queries were created for each paragraph rated by humans in the WENN dataset.

[12] http://download.wikimedia.org/enwiki/latest/

[13] PyLucene, is a Python extension that allows access to the Java version of Lucene: http://pylucene.osafoundation.org/

This WENN-based query was constructed by removing stop-words and punctuation from the title[14] that accompanied each paragraph, and then joining the remaining words with "OR" statements (see Appendix E.8). In contrast, the query devised for the paragraphs in the Lee dataset was more complex. For the Lee-based query, the researcher chose several descriptive keywords[15] for each paragraph in the Lee dataset, and used "AND" and "OR" operators to combine these keywords. Moreover, the Lee-based query used Lucene's 'star' wild-card operator to return multiple results from word stems. For example, the stem and wild-card combination "research*" would match documents containing the words "research", "researcher", and "researchers" (see Appendix E.9).

### *5.5.1. Wikipedia Sub-corpora*

Four sub-corpora were created using the Lucene queries (described above) on the Wikipedia document set. For each dataset (WENN & Lee), a 1000 document and a 10000 document sub-corpus was generated. The structure of the Wikipedia articles contained in these sub-corpora was substantially different from the documents contained in either the WENN or Toronto Star corpora (see Table E.4.1). Wikipedia articles tend to be longer in format, with documents that approximate the length of a short essay (on average 1813 to 2698 words per document). In contrast, the documents contained in both the WENN and Toronto Star corpora are similar in length to a journal article's abstract (on average 74 to 255 words per document). In addition to the Wikipedia documents being generally much longer than the WENN or Toronto Star documents, the Wikipedia documents also contain on average many more unique words.

The greater size and complexity of the Wikipedia documents may produce noise for the semantic models. However, Lee and Corlett's (2003) findings indicate that decisions about a

[14]These titles were not included with the WENN paragraphs when similarity comparisons were made by either humans or the semantic models.

[15]On average, four keywords were chosen per paragraph to form the Lee-based query.

document's content can be made using only the beginning of a document's text. In their study of Reuters' documents, words found in the first 10 percent of a document's text were judged to hold greater 'mean absolute evidence' characterizing a document's content. Lee and Corlett calculated the 'evidence' value of a word given a particular topic. This calculation was made by comparing how often a word appeared in documents related to a topic, relative to the word's frequency in documents that were not topic-related. Their finding may reflect a generally good writing style found in Reuters' documents, where articles may begin with a précis or summary of the information that is contained in the rest of the document. Documents in a Web-based medium such as Wikipedia, may also conform to this generalization. Intuitively, it seems likely that important descriptive information displayed on a Web page will be positioned nearer the top of a page (probably within the first 300 words), so as not to be over-looked by the reader as the Web page scrolls or extends beneath screen.[16]

To explore the possible effect of document length (number of words) on semantic models, corpora were constructed that contained the first 100, 200, 300 and all words from the Wikipedia sub-corpora. To illustrate this point, if the preceding paragraph was considered a document, in the first 100 word condition this document would be truncated at "...by comparing how often a word appeared in". Furthermore, to test if corpus size influenced the similarity estimates generated by the semantic models, performance was compared on the 1000 and 10000 sub-corpora for both datasets. Thus, making a 2 x 4 design (number of documents in a corpus BY number of words in each document) for each dataset. Each sub-corpus was compiled using LSA at 300 dimensions. LSA was chosen for its quick compilation speeds and because of the generally good match that has been reported between LSA and human performance on tasks comparing paragraph similarity (Lee et al., 2005; Landauer & Dumais, 1997). Moreover, in general LSA was one of the best performing models that

[16] In Web usability research and broad-sheet newspaper media terms this positioning is often referred to as being "above the fold".

incorporates a knowledge base in the previous studies presented in this paper.[17] This choice of dimensionality is supported by the findings of the first two studies in this paper, where increased dimensionality improved performance.

*5.5.1.1. Document Length.*

In general, LSA's performance was better as document length was shortened, with the best results produced by truncating documents' length at 100 words. On both datasets, LSA produced the highest correlations with the human similarity judgments using the 1000 document sub-corpora truncated at 100 words (see Figure 5.5 and Figure 5.6). This configuration produced a result (0.51) that was significantly higher than all other document number and document length combinations for the Lee dataset. On the WENN dataset, the correlation for the 1000 document corpora with documents truncated at 100 words was higher than all other cases; however, this result was not significantly higher in several cases. On both datasets, truncating documents at 100 words produced significantly higher correlations than the ALL word conditions (where document length was not truncated). These results show that improvements to model performance can be achieved by truncating documents to 100 words, and this improvement supports the earlier findings of Lee & Corlett (2003).

*5.5.1.2. Number of Documents.*

LSA's performance on both datasets was best using the smaller 1000 document sub-corpora. On the Lee dataset, when documents are truncated at 100 words, the performance of LSA is better using the 1000 document sub-corpora than the 10000 document sub-corpora ($t_{(1222)} = 4.44$, $p < 0.05$). On the WENN dataset, when documents are truncated at 100 words performance was also better for the 1000 document sub-corpora, although this difference failed to reach significance ($t_{(250)} = 1.63$, n.s.).

[17] In Study Two, LSA similarity estimates correlated 0.48 with human judgments of similarity on WENN document set.

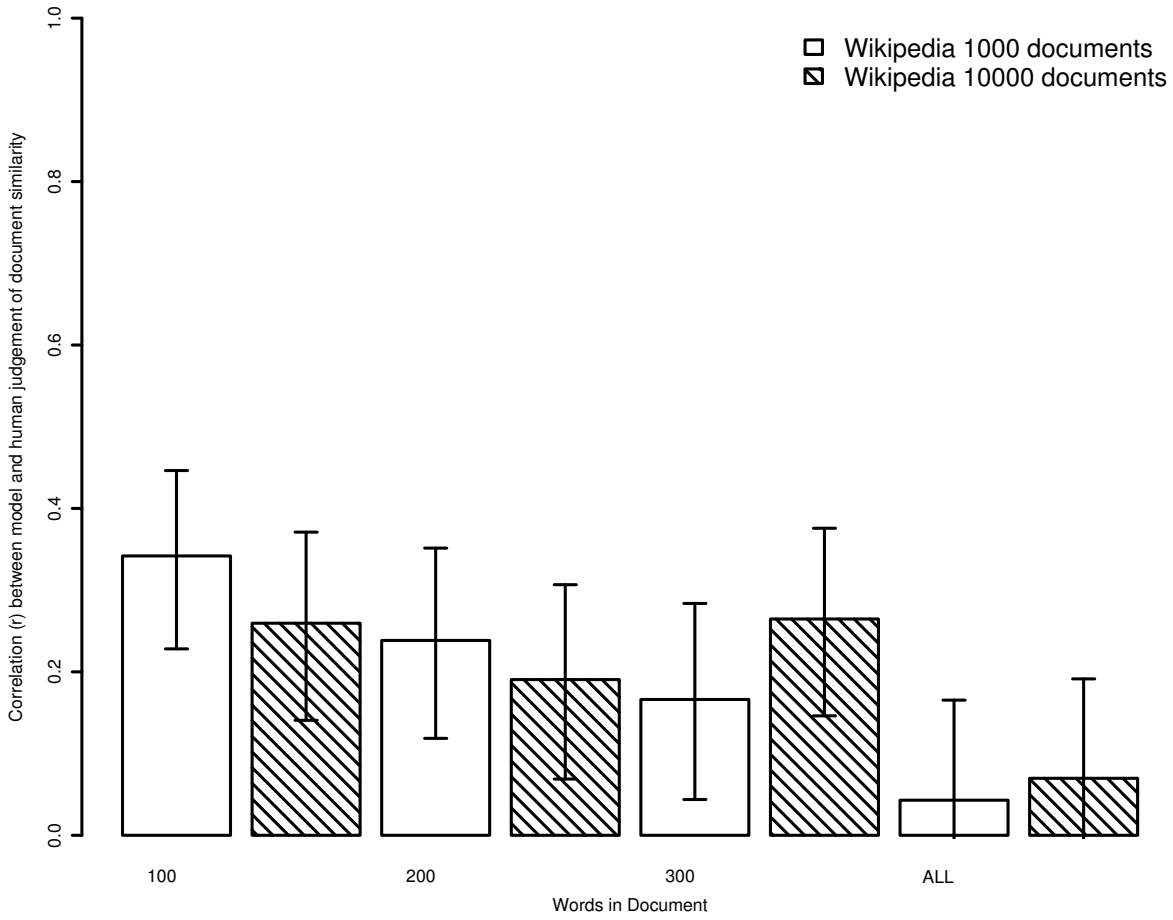



*Figure 5.5.* Correlations between human judgments of paragraph similarity on the WENN dataset with estimates made using LSA (at 300 dimensions) using the WENN Wikipedia-based corpora containing 1000 and 10000 documents retrieved using Lucene with WENN-based query. Wikipedia documents have been truncated in four ways: first 100, 200, 300, and ALL words. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

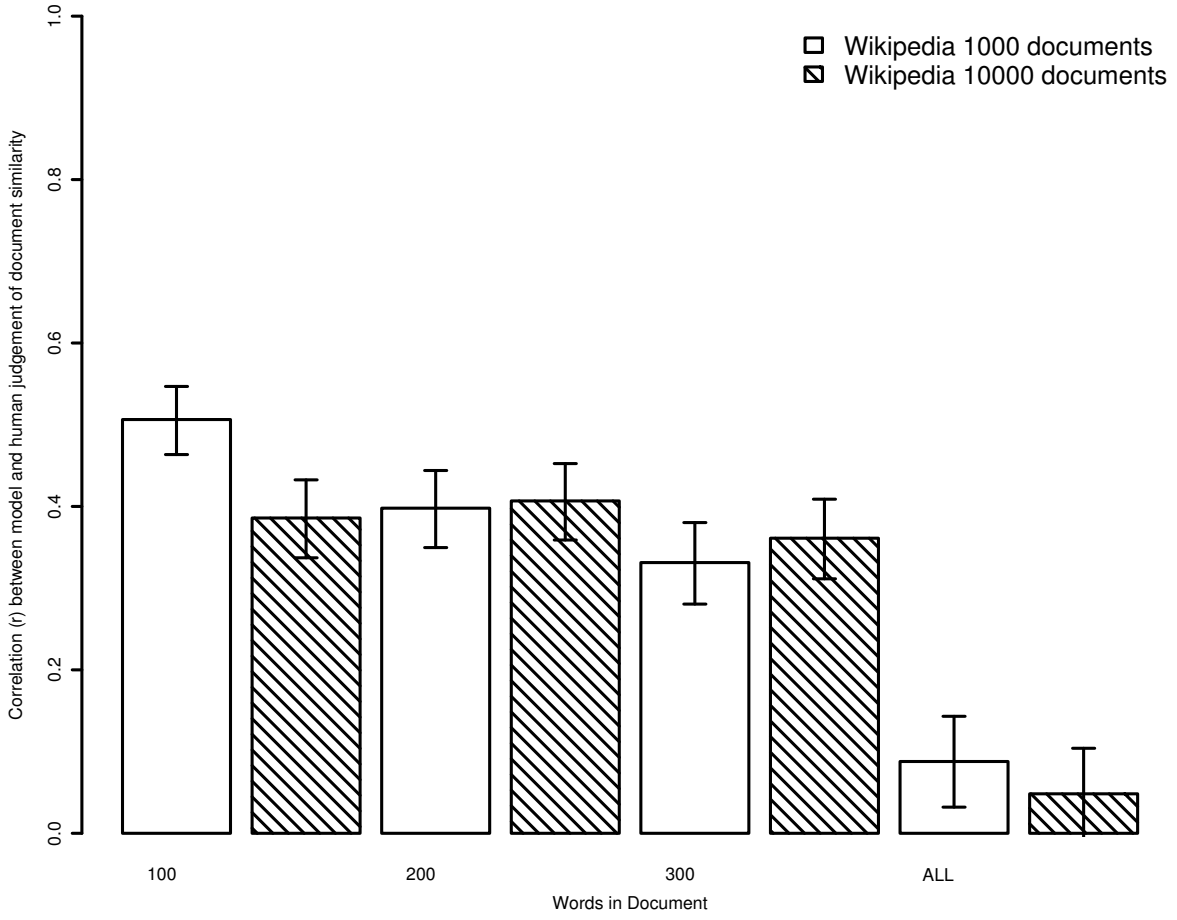



*Figure 5.6.* Correlations between human judgments of paragraph similarity on the Lee dataset with estimates made using LSA (at 300 dimensions) using Lee Wikipedia-based corpora containing 1000 and 10000 documents retrieved using Lucene with Lee-based query. Wikipedia documents have been truncated in four ways: first 100, 200, 300, and ALL words. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

*5.5.2. All models compared on Wikipedia sub-corpora*

The results presented in Study Two of this paper for models using the WENN (NN-NSL) and Toronto Star (NN-NSL) corpora have also been included in the findings presented in Figure 5.7 and Figure 5.8 as points of comparison to judge the effectiveness of creating the 1000 and 10000 document sub-corpora from Wikipedia.

When the results for both the WENN and Lee datasets are taken into consideration, again none of the more complex semantic models performed significantly better than the simple word overlap model. While the best performing model on the Lee dataset was Vectorspace (0.56) using the Wikipedia 10000 document corpus, this was not significantly different ($t_{(1222)}$ = 1.31, n.s.) from the word overlap model's correlation (0.53) with human judgments. As is displayed in Figure 5.7 and Figure 5.8, of the corpus-based models Vectorspace, LSA and SpNMF performed the best on both datasets. It is unclear whether using the Jensen-Shannon metric as opposed to dot product measure with the Topic Model produced better results. On the Lee dataset, Topic Model with dot product (0.48) using the 1000 document Wikipedia corpus significantly outperformed Topics model with the Jensen-Shannon metric (0.42) using the 10000 document Wikipedia corpus ($t_{(1222)}$ = -2.08, $p < 0.05$). However, using the WENN (NN-NSL) corpus, there was not a significant difference between the two Topic Model similarity measures ($t_{(250)}$ = 0.53, n.s.) on the WENN dataset.

LSA performed well using both the WENN (NN-NSL) and Wikipedia-based Lee corpora. Given that LSA is built on Vectorspace, it is encouraging to see that in the case of the WENN dataset dimensionality reduction improved this LSA's performance (0.48) when compared to Vectorspace (0.41). However, this improvement was not found consistently, as indicated by the higher correlation with human judgments on Lee dataset achieved by Vectorspace using either 1000 and 10000 document Wikipedia-based corpora (see Figure 5.8).

Using the WENN (NN-NSL) corpus as a knowledge base allowed the semantic models

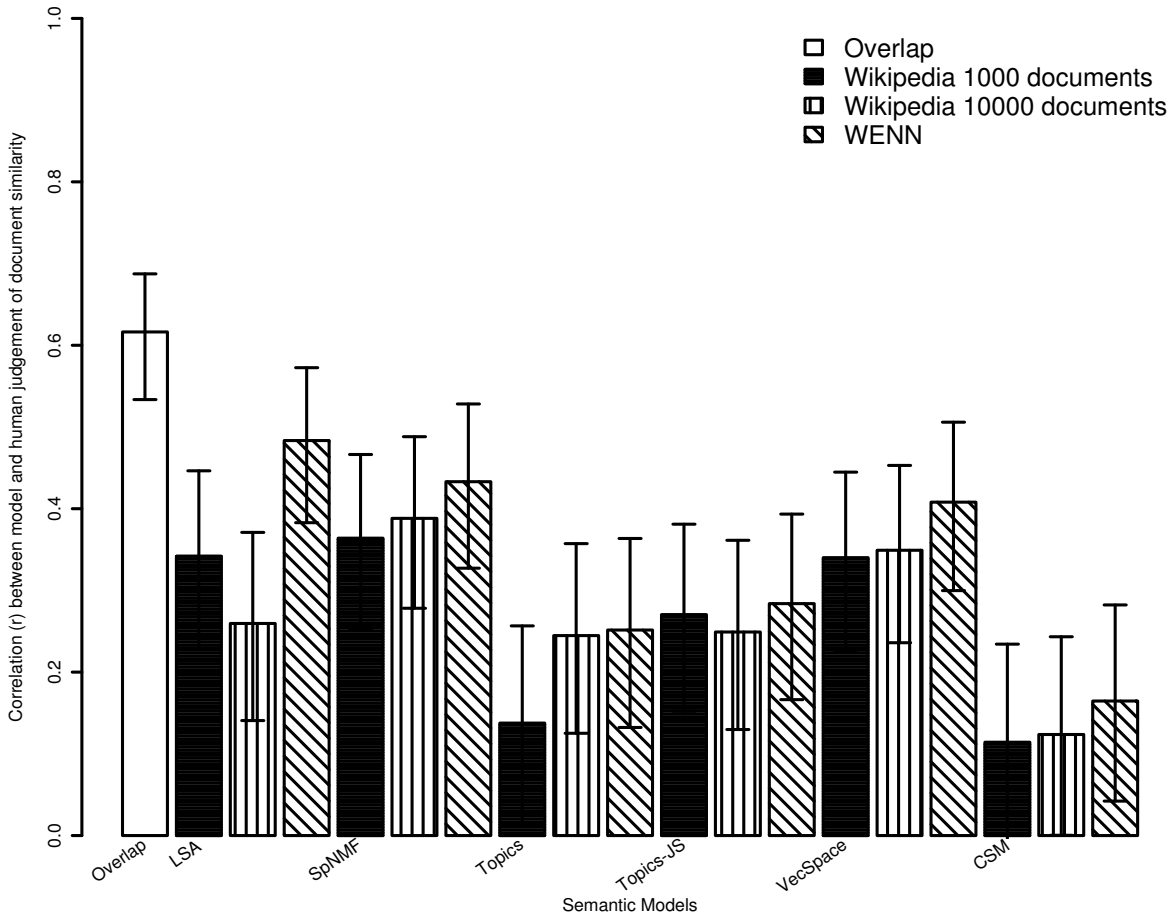



*Figure 5.7.* Correlations between human judgments of paragraph similarity on the WENN dataset with semantic model estimates made using Wikipedia Corpora with 1000 & 10000 documents and the WENN Corpus (NN-NSL). Error bars are the 95% confidence limits of the correlation. These results are also presented in Table E.10.1. Correlations exclude Same-Same paragraph comparisons.

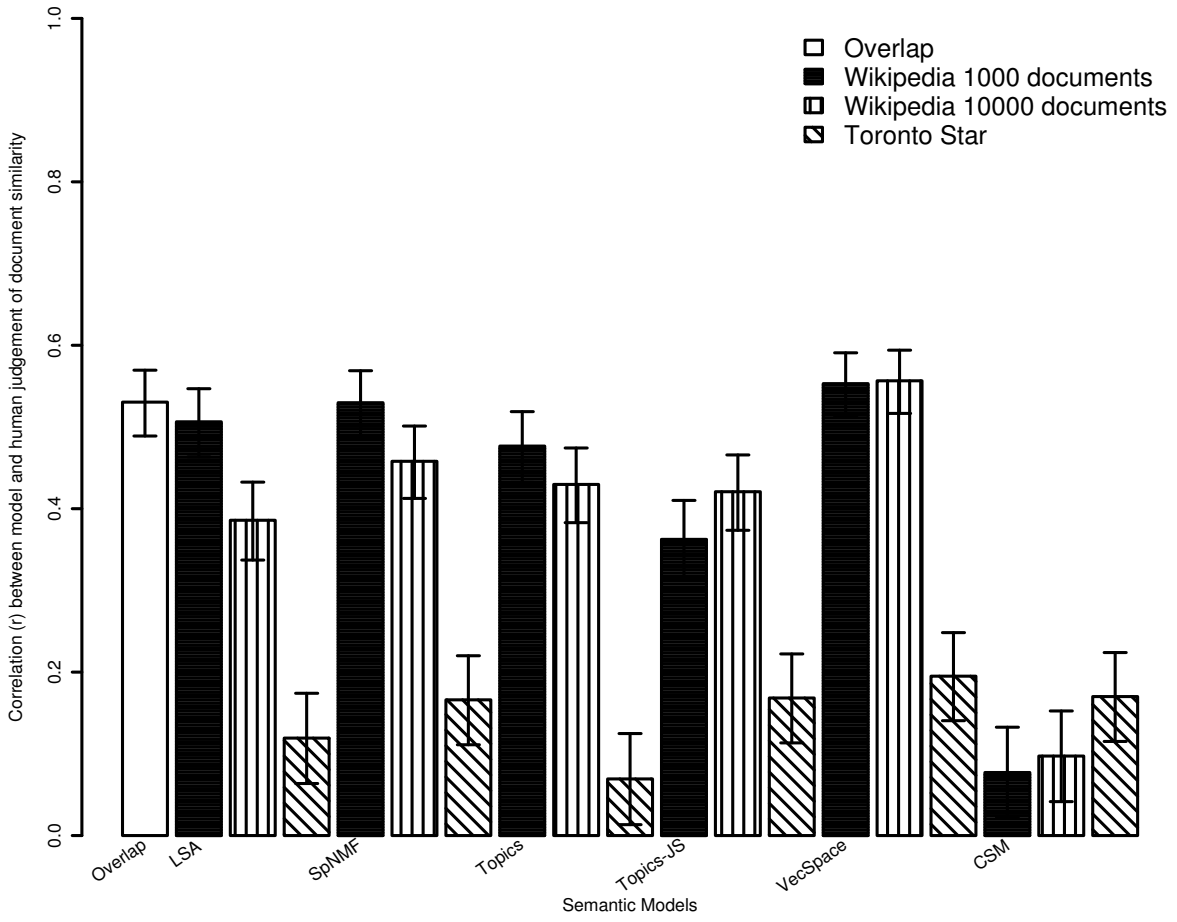



*Figure 5.8.* Correlations between human judgments of paragraph similarity on the Lee dataset with semantic model estimates made using Wikipedia Corpora with 1000 & 10000 documents and the Toronto Star (NN-NSL). Error bars are the 95% confidence limits of the correlation. These results are also presented in Table E.10.2. Correlations exclude Same-Same paragraph comparisons.

to produce better estimates of human similarity judgments than could be obtained using either 1000 or 10000 document Wikipedia-based corpora on the WENN dataset. In contrast, corpora retrieved from Wikipedia allowed the models to perform much better when making estimates of paragraph similarity on the Lee document set (see Figure 5.8). For corpus-based models, the 10000 document Wikipedia corpus was found to produce the highest correlation with human ratings on the Lee document set (Vectorspace 0.56), however in the majority of cases the 1000 document Wikipedia corpora was associated with better model performance at this task. All results presented thus far have consistently shown that the Toronto Star has provided a poor knowledge base on which to assess the paragraphs contained in Lee dataset. These results indicate that when domain-chosen corpora are not a good fit to the knowledge required to make accurate estimates of similarity on paragraphs, using corpora drawn from Wikipedia can improve model performance.

## 5.6. Study Four: Corpora that include the dataset paragraphs

In the empirical studies we have reported, subjects were presented with document pairs to be rated. Documents were repeated in different pairs, so for the majority of ratings subjects had already been exposed to all of the test documents. In the previous studies, paragraphs contained in the WENN dataset were included in the WENN corpora, but not for the corpora used by models on the Lee dataset. Consequently, the models were at a disadvantage relative to participants. This inclusion of dataset paragraphs is potentially important for models like the Topic Model where context can select for the appropriate meaning of a word. To evaluate the efficacy of including stimulus paragraphs into the semantic models' knowledge base as a method of corpus improvement, the 50 Lee dataset paragraphs were added to the most effective corpora with the most effective preprocessing found in the previous studies for the Lee dataset.

For this study, the 50 Lee paragraphs were prepended to both the 1000 and 10000 document Wikipedia corpora. These revised corpora were preprocessed using the same

techniques described in Study Three for the Wikipedia sub-corpora. While the 50 Lee paragraphs were not truncated at 100 words, preprocessing was used to remove punctuation, stop-list words, words that only appear once on the document set, numbers and single letters. After preprocessing, the smaller corpus contained 1,050 documents with 8,674 unique words and 100,107 tokens, and the larger corpus held 10,050 documents comprised of 37,989 unique words from a total of 942,696 tokens.

Adding the 50 Lee paragraphs to the Wikipedia 1000 corpora significantly improved correlations between model estimates and human judgments of similarity in nearly all cases (see Table E.11.1). While the Topics model improved one point, using the dot product measure, there was not a significant improvement using the Jensen-Shannon metric. The greatest improvement in model performance was displayed by Vectorspace which increased from 0.55 to 0.67, and LSA which rose from 0.51 to 0.60 (see Figure 5.9).

Both significant performance increases and decreases were produced for all models by prepending the 50 Lee paragraphs to the 10000 document Wikipedia corpora (see Table E.11.2). While all differences were significant when compared to non-augmented Wikipedia sub-corpora, the actual differences in performance were small for most models. In general, these differences ranged from between 0.001 to 0.02 with the exception of Topics model using the Jensen-Shannon metric which went up from 0.42 to 0.49 when the 50 Lee paragraphs were added to the Wikipedia 10000 corpus (see Figure 5.10).

## 5.7. Overall Summary

In Study One, moderate correlations were found between the word overlap model (WENN: 0.43, Lee: 0.48) and human judgments of similarity on both datasets. However, weaker performance was displayed by all of the more complex models when similarity estimates were compared on both the WENN (highest CSM, 0.26) and Lee (highest CSM 0.15) datasets. It was postulated that the semantic models' performance may have been constrained

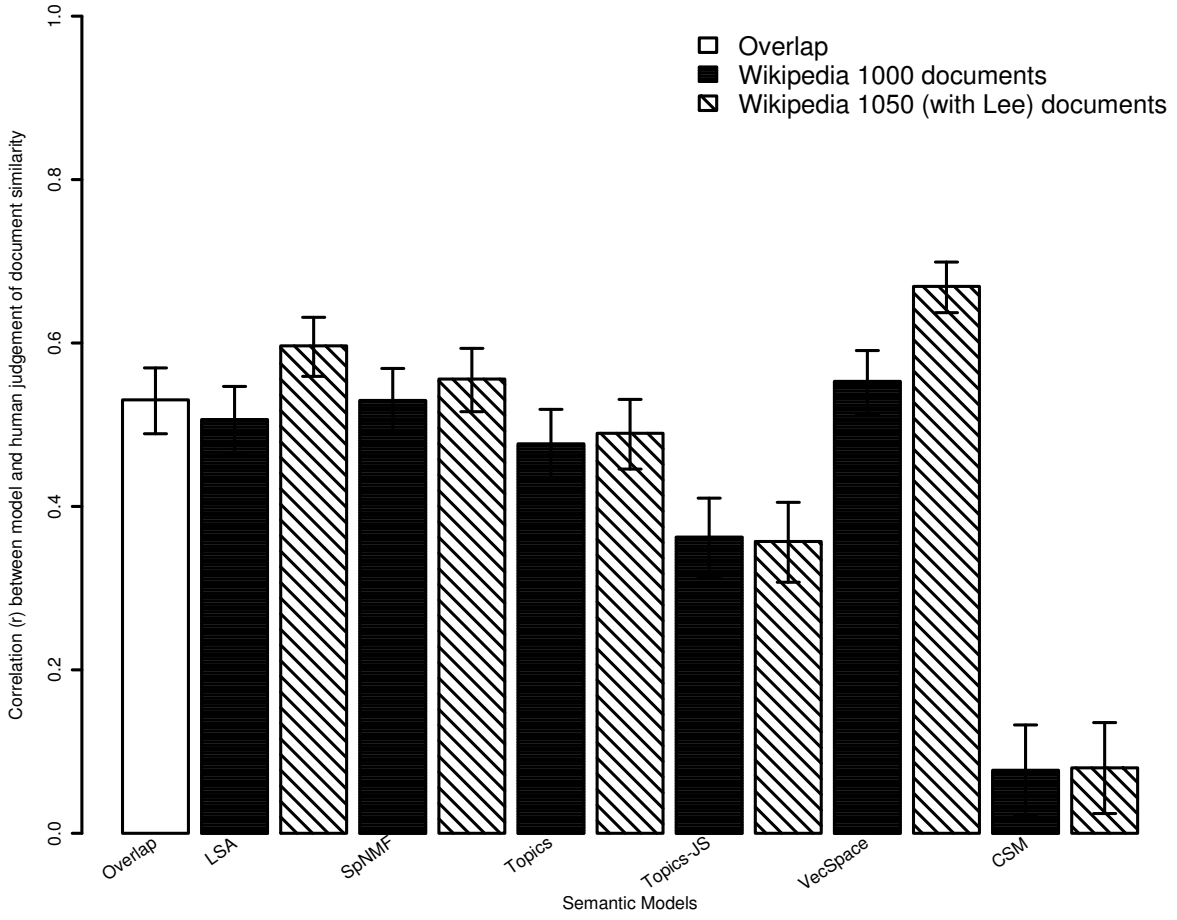



*Figure 5.9.* Correlations between human and model estimates of paragraph similarity on the Lee dataset using the standard Wikipedia 1000 corpora (Wikipedia 1000) and Wikipedia 1000 corpora including the 50 Lee documents (Wikipedia 1050). The overlap model has also been included in this bar graph to allow the reader another point of comparison. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

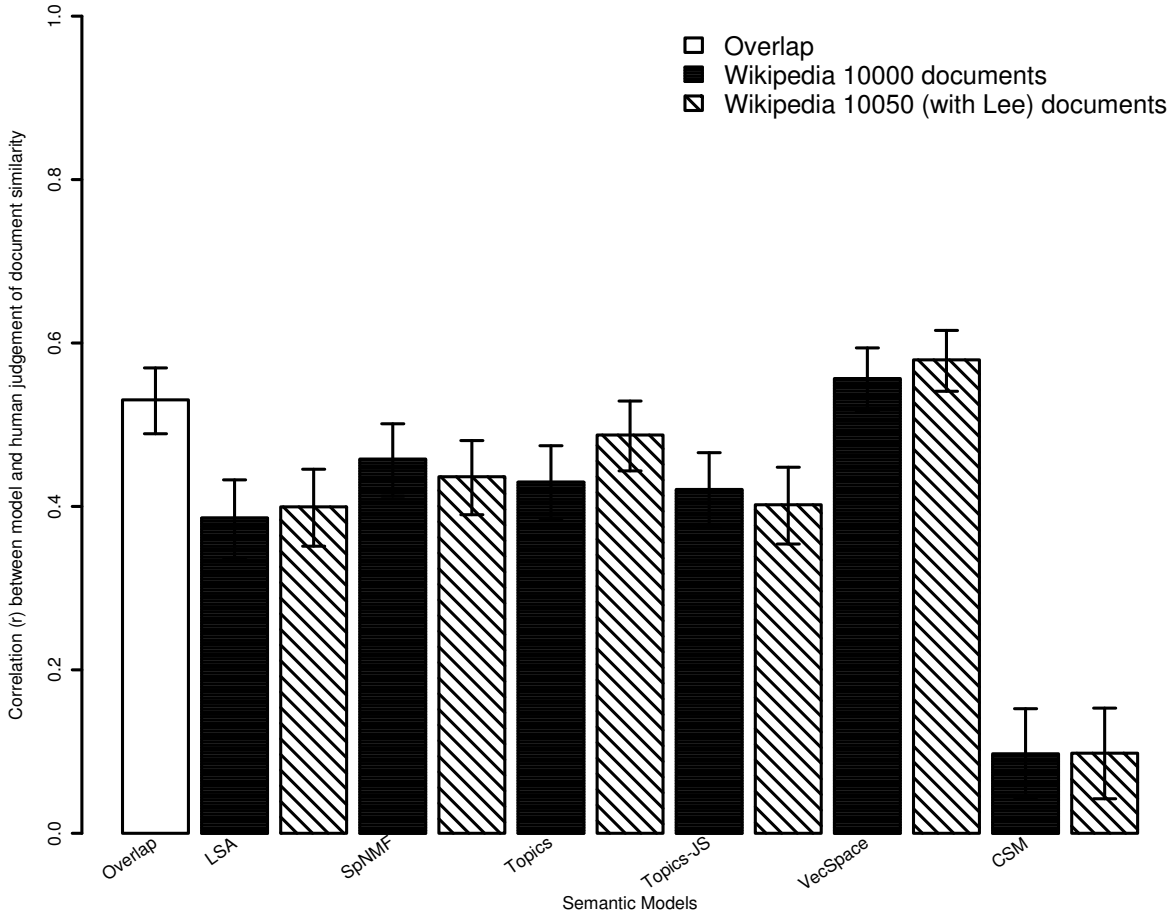



*Figure 5.10.* Correlations between human and model estimates of paragraph similarity on the Lee dataset using the standard Wikipedia 10000 corpora (Wikipedia 10000) and Wikipedia 10000 corpora including the 50 Lee documents (Wikipedia 10050). The overlap model has also been included in this bar graph to allow the reader another point of comparison. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

by factors such as corpus preprocessing and a poorly represented knowledge domain (in the case of the Toronto Star corpus and the Lee dataset). In Study Two, the importance of corpus preprocessing was highlighted, removing the numbers and single letters from corpora improved correlations with human judgment on both datasets for all models with the exception of CSM. After the removal of these characters, the best performing of the more complex models were LSA (0.48) on the WENN dataset and Vectorspace (0.20) on the Lee dataset. However, the corpus-based models still failed to outperform the word overlap model, which also improved with the removal of numbers and single letters on both datasets (WENN: 0.62, Lee: 0.53).

In some ways it is unsurprising that the models' performance in this study was better on the WENN dataset than the Lee dataset, because the paragraphs used in similarity judgments were drawn from the greater set of documents contained in the WENN corpus. That is, in the case of the WENN set there was a better match between paragraphs that were compared (WENN dataset) and the models' knowledge base (the WENN Corpus). Conversely, the Toronto Star articles did not provide a good approximation of the knowledge required to make reliable inferences regarding the similarity of paragraphs contained in the Lee dataset. While the Toronto Star corpus contains extracts of current affairs, these articles (published in 2005) must vary substantially from the précis published in 2001 that are contained in the ABC news mail service that was used by Lee et al. (2005).

In an attempt to obtain a better representation of the knowledge base required to make accurate paragraph similarity comparisons, in Study Three Wikipedia sub-corpora were generated to use on each dataset. The Wikipedia documents were found to be much longer and more like short essays than the summary or abstract length documents found in the WENN and Toronto Star corpora. Guided by the research findings of Lee and Corlett (2003), it was found that Wikipedia documents truncated at 100 words provided better corpora for LSA at 300 dimensions than when using all of the words contained in these documents. LSA's

performance was also better using the smaller 1000 document Wikipedia sub-corpora. The decision to use 300 dimensions was in part based on the results of Study One and Study Two, which indicated that increased dimensionality often led to significant performance gains when model estimates of paragraph similarity were compared to human ratings.

Based on these findings, spaces were compiled for the models using Wikipedia corpora that contained documents truncated at 100 words. The semantic models' performance on the WENN dataset did not improve using these Wikipedia sub-corpora when compared to results achieved by models using the WENN corpus. However, there was a substantial improvement by nearly all models (except CSM) when similarity estimates were compared on the Lee dataset. Using the Wikipedia sub-corpora, the best performing of the more complex models on the Lee dataset were Vectorspace using both 1000 documents (0.55) and 10000 documents (0.56) and SpNMF using 1000 documents (0.53); all of which approach the inter-rater reliability (0.6) recorded for Lee and colleagues participants (Lee et al., 2005). The decrement in performance seen using the Wikipedia sub-corpora, when compared to the WENN corpus on the WENN dataset, is again somewhat expected given that the documents in the WENN dataset were selected from the WENN corpus. When the results on both the WENN and Lee datasets are considered, Vectorspace, LSA and SpNMF were the best performing of the corpus-based models. That said, even using corpora that allowed models to perform on a comparable level with the inter-rater reliability found in the WENN dataset, and approaching that calculated for the Lee dataset, these models still could not significantly outperform the simple word overlap model when estimating the similarity of paragraphs in comparison to human performance at this task.

The final study explored what effect including the dataset paragraphs into a corpus would have on models' performance. This assessment was only undertaken for the Wikipedia corpora used on the Lee dataset, as the WENN documents were already included in the WENN corpora in previous studies. In particular, the Topic Model performance was expected to

increase; however, this improvement in performance was only observed for the Topic Model using the Dot Product measure of similarity. Generally, performance increases associated with the inclusion of the 50 Lee paragraphs were greater on the smaller 1050 document Wikipedia corpus when compared to those observed on the 10050 Wikipedia corpus. It is possible that any benefit to a model's performance produced by adding these 50 paragraphs is negated by the volume of terms contained in the larger corpus. Overall, the best performance was observed for Vectorspace (0.67) and LSA (0.60) using the 1050 document Wikipedia corpus containing the 50 Lee paragraphs. It was interesting to note, that LSA's performance using the smaller Wikipedia corpus and 50 Lee paragraphs was almost exactly the same as the inter-rater reliability calculated for the Lee dataset. Furthermore, using this augmented 1050 document Wikipedia corpus, both LSA ($t_{(1222)} = 3.20$, $p < 0.05$) and Vectorspace ($t_{(1222)} = 7.81$, $p < 0.05$) significantly outperformed the overlap model (0.53) when estimates of paragraph similarity were compared to the human judgments contained in the Lee dataset.

Figure 5.11 displays scatterplots from the two best performing models on WENN and Lee datasets. It was surprising that on both datasets, the simple word overlap model was among the two best performing models. As is illustrated by Figures 5.11B and D, the word overlap model is generally capturing human responses that have been made on paragraphs which have low or no similarity. It is also interesting to note that on the WENN dataset, LSA using the WENN corpus (NN-NSL) has in all cases estimated some similarity between the paragraph pairs (see Figure 5.11A). This may indicate that greater dimensionality is needed by LSA to adequately delineate the themes presented in the WENN corpus documents. At another level, because the WENN paragraphs all focus on "celebrity gossip", to some extent they may all be considered related. Alternatively, on the Lee dataset, Vectorspace appears to have provided a relatively good match to the average human estimates of paragraph similarity (see Figure 5.11C).

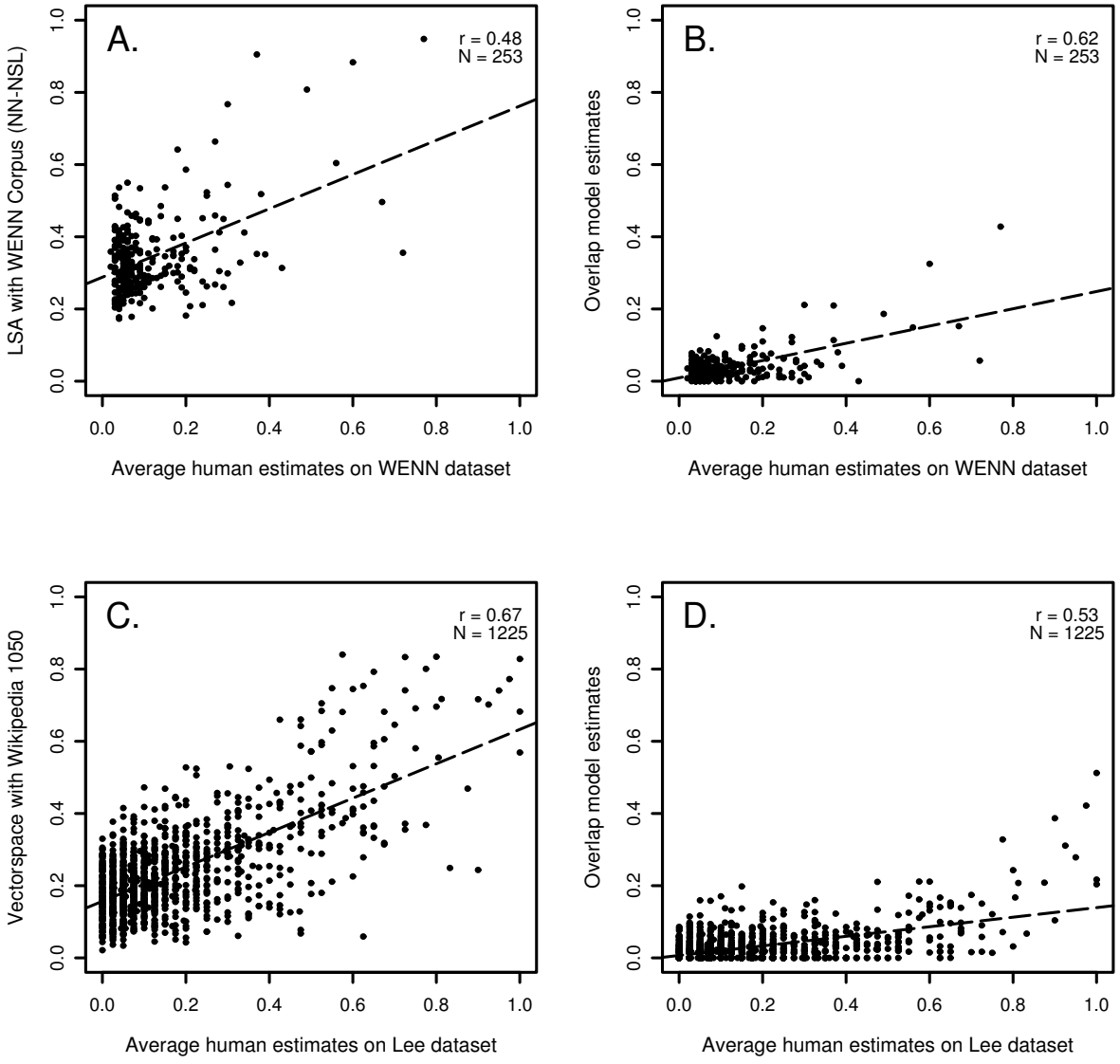



*Figure 5.11.* Scatterplots of the two best similarity estimates calculated for both the WENN and Lee datasets compared to the average similarity estimates made by humans for each pair of paragraphs. On the WENN dataset, (A) LSA using the WENN corpus (NN-NSL), and (B) the Overlap model. On the Lee dataset, (C) Vectorspace using the Wikipedia 1050 (including Lee documents), and (D) the Overlap model. Note, on the Lee dataset, average human ratings have been normalized [0,1].

## 5.8. Discussion

Quite surprisingly, the simplest models (Vectorspace and word overlap) were the best performing models examined in this research, both exceeding the inter-rater reliability calculated for human judgments. While exceeding the inter-rater reliability is an important milestone, it is possible for a model to perform better. The model is compared against the average rating of the subjects, which eliminates a significant amount of variance in the estimates of the paragraph similarities, whereas the inter-rater reliability is the average of the pairwise correlation of the subjects. On the WENN dataset, the overlap model (0.62) exceeded the inter-rater reliability (0.47). Similarly the Vectorspace model (0.67) using a corpus containing both truncated Wikipedia documents and the 50 Lee paragraphs also exceeded the inter-rater reliability found for the Lee dataset (0.605).

The Vectorspace model's performance on the Lee dataset using the smaller Wikipedia corpus that included the 50 Lee paragraphs was particularly encouraging. While the overlap model's good performance at these tasks can largely be accounted for by its ability to capture human ratings on paragraph pairs with low or no similarity, the Vectorspace model appeared to provide good estimates of both the similarity and dissimilarity of the Lee paragraphs when compared to human ratings. That said, the Vectorspace model did not perform as well on the WENN dataset, when compared to estimates produced by either the overlap model or LSA. However, the finding that no model performed as well as the overlap model on this dataset, might indicate that even though the best results for the WENN dataset were found for most corpus-based models using the WENN corpus (NN-NSL), that this corpus still did not provide an adequate term representation for the models. Furthermore, it is possible that a better match to the background knowledge needed by models for the WENN paragraphs may have been accomplished had a more complex Lucene query been used to retrieve relevant Wikipedia documents.

One possible explanation for the success of the overlap and Vectorspace models in these studies may be found in the framework of the experiments. In each experiment, participants made pairwise comparisons of paragraphs displayed on a computer monitor. The side-by-side positioning of these paragraph pairs may have encouraged keyword-matching (or discrimination) between the paragraphs by the participants. That is, it is possible that the participants were skimming paragraphs for keywords with which they could make similarity judgments. Another related strategy which could result in the similar outcome, would be to read one paragraph thoroughly and then to skim the comparison paragraph for salient words presented in the first text. Masson (1982) indicates that when skimming, readers miss important details in newswire texts, and that visually unique features of text such as place names may increase efficiency of skimming as a reading strategy. Given that names of people and places were certainly present in all paragraphs presented to participants in this research, commonalities between participants' similarity estimates (and also those of the overlap and Vectorspace models) may also be influenced by these proper nouns. In future research, eye-tracking technology could be employed to elucidate the possibility of skimming strategies in this type of experimental task. Alternatively, paragraphs could be presented in a serial sequence, rather than concurrently, and time spent reading each paragraph might act as an indicator of reading strategy.

In the introduction, we categorized the materials used in this class of research as having four types of textual unit: words, sentences, single paragraphs, and chapters or whole documents. Past research has indicated that the Topic Model performs better at word association tasks than LSA. Moreover, researchers have shown that Topic Model's ability to accommodate homographs is superior to other models at the single word textual unit level (Griffiths et al., 2007). While the ability to discriminate the intended meaning of ambiguous words is certainly desirable, it is possible that this attribute is not a prerequisite for successful model performance with larger textual units such as paragraphs. This may be because textual

units such as sentences and paragraphs allow models access to a range of non-ambiguous words whose informativeness may compensate a model's inability to capture the meaning of more ambiguous words (Landauer & Dumais, 1997; Choueka & Lusignan, 1985). In the studies reported above, four models (word overlap, Vectorspace, LSA, and SpNMF) that do not capture this type of word ambiguity all outperformed the Topic Model when compared to human ratings at the task of estimating similarity between paragraphs.

Besides a model's ability to make good approximations at human similarity judgments, another factor that must be considered when evaluating the usefulness of these semantic models is the ability to produce interpretable dimensions. For example, one of the criticisms of LSA is that the dimensions it creates are not always interpretable (Griffiths et al., 2007). Similarly, word overlap, Vectorspace and CSM do not employ any dimensionality reduction, and thus provide word vectors that are difficult to interpret. In contrast, both SpNMF and Topic Model return interpretable dimensions. To illustrate this point, Table E.10.3 displays a sample of the dimensions created for the 10000 document Lee-based Wikipedia corpus. As is made clear in this table, it would be easy to meaningfully label any of these dimensions.

Given its generally good approximations of human judgment and ability to provide interpretable dimensions, SpNMF could be regarded as one of the best models examined in this article. However, its slow compilation of spaces would certainly need to be addressed for it to be generally useful in either a research or an applied setting. In comparison to the Vectorspace model which takes 24 seconds to compile a space of the King James Bible using a 2.4 GHz CPU, the SpNMF model is very computationally expensive taking just under 8 hours. Future research may be able to utilize parallel programming techniques[18] to sequence SpNMF calculations over multiple processing units to reduce the time needed to compile SpNMF spaces, and thus make SpNMF a more feasible model to use in tasks of this kind.

[18]CUDA, the nVidia graphics processing unit technology, presents as an architecture on which these parallel processing gains might be achieved whilst efficiently using sparse matrices (Bell & Garland, 2008).

In several ways, CSM was the worst performing model employed in this research. All models performed better than CSM when using either the Wikipedia sub-corpora or WENN corpus (NN-NSL). Also, the matrices contained within CSM spaces can be over two orders of magnitude larger than those compiled by other models. For example, the space compiled by CSM for the 10000 document Wikipedia-based corpus with documents truncated at 100 words for the Lee dataset was 12Gb in size. In stark contrast, the same corpus compiled by Vectorspace used 84Mb of disk space. While files of this size are not unusable, accessing the dense vectors contained in CSM spaces is slower than retrieving vectors for comparisons using the other models.

One of the key strengths of the simple overlap model that performed so well in this research, is that it is not reliant on a knowledge base, only extracting information from the surface structure of the textual stimuli. This paper has provided examples of the difficulties researchers face when attempting to create a suitable knowledge base for semantic models. This is not to mention the labor intensive process undertaken to collect and format a large corpus. Furthermore, the simple overlap model is not without theoretical underpinning. Max Louwerse, in this issue of TopiCS, has suggested that "support for language comprehension and language production is vested neither in the brain of the language user, its computational processes, nor embodied representations, but outside the brain, in language itself". In arguing his claim, Louwerse provides examples of how first-order co-occurrences of terms can produce similar results to LSA on tasks of categorization. Similarly, it could certainly be argued that to some extent the good performance of the overlap model in the studies presented in this paper support Louwerse's argument.

Overall, dimensionality reduction did not appear to improve the models' estimate of paragraph similarity when compared to results produced by Vectorspace and overlap models. However, LSA's consistent performance, mimicking of human inter-rater reliability, and better performance on the WENN dataset when compared to Vectorspace, all indicate that

there is further research that must be done in this area. One aspect of this research that we intend to explore more fully, is the possibility that subsets of participants' judgment variance can be accommodated by different models. For example, it is possible that participants who are skimming the paragraphs may produce results more similar to either Vectorspace or the overlap model. In contrast, other participants who are reading carefully and not skimming over the text, may produce results that are more similar to those calculated with LSA. While it is not possible to draw these conclusions with any certainty based on our current datasets, eye-tracking technology will be employed in future research to explore these possibilities.

The findings presented in this paper indicate that corpus preprocessing, document length and content are all important factors that determine a semantic model's ability to estimate human similarity judgments on paragraphs. The online, community driven Wikipedia encyclopedia also proved to be a valuable resource from which corpora could be derived when a more suitable domain-chosen corpus is not available. In many applications the hand construction of corpora for a particular domain is not feasible, and so the ability to show a good match between human similarity judgments and machine evaluations is a result of applied significance.

# Chapter 6. Semantic Models and Corpora Choice when using Semantic Fields to Predict Eye Movement on Web pages (submitted)

Benjamin Stone

School of Psychology, The University of Adelaide

Simon Dennis

Department of Psychology, Ohio State University

submitted, International Journal of Human-Computer Studies.

Statement of Contributions

Benjamin Stone (Candidate)

I was responsible for the conception and primary authorship of the paper. I was responsible for the development and programming of software based assessment tools, and the collection and modeling of all data. I conducted the statistical analyzes independently with advice from the co-author. I am the corresponding author and primarily responsible for responses to reviewers and revisions to the paper.

SIGNED: DATE: 10/8/10

Simon Dennis (Co-author)

(see Appendix D)

## 6.0. Abstract

Ten models are compared in their ability to predict eye-tracking data that was collected from 49 participants' goal-oriented search tasks on a total of 1809 Web pages. Forming the basis of six of these models, three semantic models and two corpus types are compared as components for the Semantic Fields model (Stone & Dennis, 2007) that estimates the semantic salience of different areas displayed on Web pages. Latent Semantic Analysis, Sparse Non-Negative Matrix Factorization, and Vectorspace were used to generate similarity comparisons of goal and Web page text in the semantic component of the Semantic Fields model. Surprisingly, Vectorspace was consistently the best performing semantic model in this study. Two types of corpora or knowledge-bases were used to inform the semantic models, the well known TASA corpus and other corpora that were constructed from the Wikipedia encyclopedia. In all cases the Wikipedia corpora outperformed the TASA corpora. A non-corpus based Semantic Fields model that incorporated word overlap performed more poorly at these tasks. Three baseline models were also included as a point of comparison to evaluate the effectiveness of the Semantic Fields models. In all cases the corpus-based Semantic Fields models outperformed the baseline models when predicting the participants' eye-tracking data. Both final destination pages and pupil data (dilation) indicated that participants' were actively performing goal-oriented search tasks.

## 6.1. Introduction

The exponential increase in Internet usage over the last decade has motivated psychological researchers to examine Web users' behavior in this virtual environment. Research focusing on user behavior in Web page environments can generally be delineated into two main streams: display-based and semantics-based research. While both methods to some degree attempt to predict the area on a Web page that a user will focus their attention towards, they approach this task in different ways. Display-based research has focused on perceptual aspects of the Web page, and explores components such as element and menu position, color usage, font style, and animation of graphics (Faraday, 2000, 2001; Ling & Van Schaik, 2002, 2004; McCarthy et al., 2003; Pearson & Van Schaik, 2003; Grier, 2004; Rigutti & Gerbino, 2004). Alternatively, when attempting to predict users' Web page navigation, semantic-based research matches Web users' information needs to the concepts displayed within the textual content of Web pages (Blackmon et al., 2002; Chi et al., 2003; Pirolli & Fu, 2003; Brumby & Howes, 2003, 2004; Cox & Young, 2004; Kaur & Hornof, 2005). Overall, display-based and semantics-based research into Web users' visual search of Web page hyperlinks has indicated that the user's search processes are influenced by factors such as: text semantics, element position, aesthetic qualities of elements, and environmental learning.[1]

Several researchers have highlighted the importance of combining display-based and semantic information when modeling users' navigation through Web sites (Blackmon et al., 2002; Chi et al., 2003; Pirolli & Fu, 2003; Kaur & Hornof, 2005; Stone & Dennis, 2007). Research that has combined display and semantic information when predicting Web users' behavior include the Cognitive Walkthrough for the Web (CWW, Blackmon et al., 2005), the Bloodhound Project (Chi et al., 2003), and the Latent Semantic Analysis - Semantic Fields model (LSA-SF, Stone & Dennis, 2007). For a detailed description of the CWW and

[1] For Web site navigation and environmental learning see Pan et al. (2004).

Bloodhound Project the reader is directed to Blackmon et al. (2005) and Chi et al. (2003), respectively.

### *6.1.1. Semantic Fields (SF)*

In a previous article, we presented the LSA Semantic Fields (LSA-SF) model (Stone & Dennis, 2007), which was used to predict the eye movements of 49 participants recorded during goal-oriented search tasks on three Web sites. The LSA-SF model used Latent Semantic Analysis (LSA, Landauer et al., 2007) to calculate the similarity between a textual representation of the users' goal and each of the textual elements displayed on a Web page. Using a decay function, these LSA estimates of similarity were then distributed and summed over each pixel position for all of the textual elements contained on a Web page (see Equation 1). Combining the semantic information ($L$) with distance ($d_{i(x,y)}$) from its display position using a decay function, enabled the production of maps of information density for each Web page in our study (see Figure 6.1).

$$SF_(x,y) = \sum_i L_i e^{-\lambda d_{i(x,y)}} \quad (1)$$

### *6.1.2. Focus of this paper*

Initially, the TASA corpus was used as a knowledge based for the LSA component of the LSA-SF model (Stone & Dennis, 2007). However, recent research has suggested that a better fit with the knowledge domains required to model human similarity judgments in document comparison tasks may be achieved using document sets which are retrieved from the online encyclopedia Wikipedia (Gabrilovich & Markovitch, 2007; Stone, Dennis, & Kwantes, in press). Also, while LSA is certainly the best studied statistical semantic model, Stone et al. (in press) found that other models such as the vector space model (Vectorspace, Salton et al., 1975) and Sparse Nonnegative Matrix Factorization (SpNMF, Xu et al., 2003; Shashua &

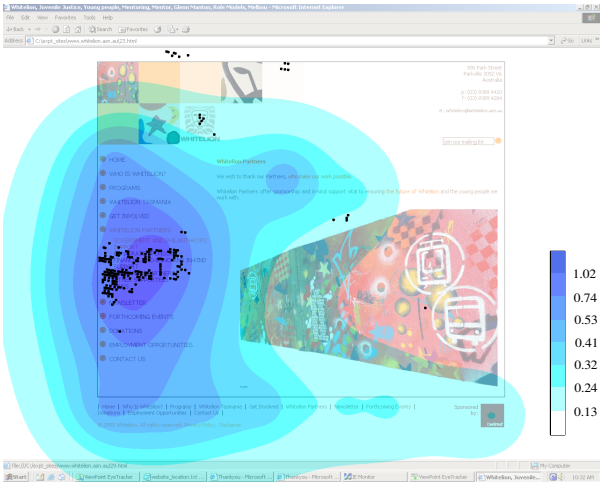



*Figure 6.1.* Semantic Fields Map using Vectorspace and a corpus drawn from Wikipedia. Participant's eye tracking data is super imposed using black dots. While the original SF model only used LSA, the SF models presented in this paper incorporate word overlap, Vectorspace, LSA, and SpNMF semantic models. Areas of greater estimated goal-oriented information salience have darker colors in this heat map.

Hazan, 2005) outperformed LSA when estimating human judgments of paragraph similarity. Based on these findings, in this paper we present a comparison of three semantic models (LSA, SpNMF, and Vectorspace), and two types of knowledge base (TASA and Wikipedia), when used as components in the generation of Semantic Fields. Furthermore, the performance of these revised Semantic Fields models is assessed on the eye-movement dataset presented in Stone and Dennis (2007).

## 6.2. Method

### *6.2.1. Participants*

Eye-movement data generated by 49 university participants, 27 males and 22 females, on three Web sites was recorded during nine goal-oriented search tasks. Participants were

recruited from either a first-year pool volunteering in exchange for partial course credit, or other members of Adelaide University who were paid $30 for their time. Participant ages ranged from 16 to 57 (M=22y3m SD=1y). Also, there was a positive skew in the samples age distribution, with 93 percent of participants were younger than 31 years old.

All participants had both previous computer and Internet experience. Self-reported years of Internet experience ranged from 4 to 17 years (M=8y8m, SD=4m), with self-reported frequency of Internet use ranging between 2 to 75 hours per week (M=14h14m, SD=1h52m). Based on these self reports, the group appeared on average to be very experienced users of the Internet.

### *6.2.2. Apparatus*

#### *6.2.2.1. Behavioral recording equipment and software.*

The IETracker program was developed to record both participants' behavior during Web site search tasks and Web page display characteristics. These observations are accomplished by programmatically controlling, and integrating, Microsoft Internet Explorer (Version 6) and the ViewPoint EyeTracker PC-60 QuickClamp System. User and site specific data collected during this exploratory experiment included: eye-tracking; hyperlink clicking; and Web page composition (location, semantics, images, color, style, and size of Web page elements). All data was then stored in a Postgresql database for later analysis.

#### *6.2.2.2. Web sites.*

Three Web sites were chosen from the Internet:

www.missionaustralia.com.au (Mission Australia)

www.greencorps.com.au[2] (Green Corps Australia)

www.whitelion.asn.au (White Lion Australia)

[2]The Green Corps Australia URL is no longer used, the Australian Government has changed both the Web site and its URL, which can be viewed here: http://www.greencorps.gov.au

Static versions of these sites[3] were pre-fetched in December 2005 to avoid changes created by Web site updates. These Web sites are all similar in the type of service they provide, such that they all offer services to disadvantaged members of the community. The Web sites were chosen because they were sufficiently complex that searching for information on these sites would be a non-trivial task for participants.

The original Stone and Dennis (2007) study used eye-tracking data recorded on 1842 Web pages, in this study only 1809 Web pages are used. It was found that 33 page views included in the original dataset had to be removed for two reasons. Some pages were omitted because data for all calibration points had not been recorded. This stemmed from participant head movements and overt glances during the calibration procedure. Also, several Web pages had been designed to catch user clicks on PDF files. These "catch-pages" only contained one textual element that informed the participants that they had either found their goal or had not and should click the "back" button. These pages have been removed because they were not part of the original Web sites, and their simple one element construction with black text on a white background probably favored the Semantic Fields model.

### *6.2.3. Procedure*

Participants were required to search for the following three target pieces of information on each of the three Web sites.

Mission Australia:

1. Who is currently the Chief Operating Officer of Mission Australia?

2. You are interested in working for Mission Australia. Search their Web site for the current job vacancies available at Mission Australia.

3. You are currently researching homelessness in young people and have heard that Mission Australia has recently published a report called "The voices of homeless young

[3]The static versions of these Web sites can be found here: http://www.psychology.adelaide.edu.au/mall_lab/lsa-sf_sites/

Australians". Search the Mission Australia Web site for this report into youth homelessness.

Green Corps:

1. You want to know more about Green Corps management. Find out who is the National Program Manager of Green Corps.

2. Find what environmental and heritage benefits are contributed by Green Corps.

3. Find the online Expression of Interest form to apply to become a Green Corps Partner Agency.

White Lion:

1. Find out who is the current President of White Lion.

2. You are interested in becoming a mentor for young people. Find out how to become one of White Lions mentors.

3. You are interested in financial viability of White Lion as a business. Find out which Government Departments are supporters of the White Lion organization.

Each task was read aloud to the participants twice before they commenced their search. Moreover, they were asked to signal the experimenter (with a hand gesture) at any time they wanted the search task repeated aloud. A three by three Latin square design was used to control for order effects in the display of each Web site. Also, a nested three by three Latin square design was used for the same purpose to guide the presentation order of each of the target tasks.

After an initial calibration procedure, using the Viewpoint eye tracking software, participants were given their search task in the manner described above and commenced their search. Given the physical structure of the Viewpoint Eye-Tracker (which uses a mounted camera with forehead and chin rests) and some participant's predisposition to fidget, it was necessary to perform additional eye-tracking calibration during the search tasks. This additional calibration required the participants to focus on targets in nine separate regions of the monitor. Moreover, these targets were automatically displayed on the screen by the

IETracker program after each hyperlink clicked during the participant's search task. To elaborate on this point further, if a participant clicked through four pages in their search for the target information, then four extra calibration procedures were performed during this search task; each calibration performed after leaving the page that was clicked on and before the next page was displayed to the participant.

The calibration of the eye-tracking data was performed in a different way in this study compared to Stone and Dennis (2007). In the previous study, an algorithm was used to adjust the eye-positions relative to participants' movement during experimentation. However, to make this process more transparent in the current study, the participants' eye-points were repositioned using the average offsets recorded over all nine calibration points.

Finally, participants were instructed that when they believed that they had found the target information that they should then click on the 'HOME' icon (which is an image of a house on Internet Explorer's menu bar). This was followed by one more round of automated calibration. On average, it took participants three page views to find their goal on the White Lion Web site, four page views on the Green Corps Web site, and five page views on the Mission Australia Web site.

### *6.2.4. Semantic Fields Models*

Four semantic models were incorporated into the Semantic Fields model[4]: word overlap, Vectorspace model (Salton et al., 1975), Latent Semantic Analysis (Landauer et al., 2007), and Sparse Nonnegative Matrix Factorization (Xu et al., 2003).

#### *6.2.4.1. Word Overlap.*

Simple word overlap was used as a baseline in this research, and also because of its solid

[4]The SEMMOD semantic models package was used to incorporate the Vectorspace model, Latent Semantic Analysis, and Sparse Nonnegative Matrix Factorization into the Semantic Fields model. The SEMMOD semantic models package is released under the GNU License and can be found here: http://mall.psy.ohio-state.edu/wiki/index.php/Semantic_Models_Package_%28SEMMOD%29

performance at paragraph level comparisons (Stone et al., in press). Term frequencies are calculated for each document, and similarities are then measured as cosines (see Equation 2) of the resulting document vectors.

$$\cos\theta = \frac{v_1 \cdot v_2}{\|v_1\| \, \|v_2\|} \tag{2}$$

*6.2.4.2. The Vectorspace model.*

The Vectorspace model (Salton et al., 1975) assumes that terms can be represented by the set of documents in which they appear. Two terms will be similar to the extent that their document sets overlap. To construct a representation of a document, the vectors corresponding to the unique terms are multiplied by the log of the frequency within the document and divided by their entropy across documents and then added. Using the log of the term frequency (TF) within documents identifies higher frequency or important words in those documents. While dividing by the entropy reduces the impact of high frequency words that appear in many documents in a corpus. Similarities are measured as the cosines between the resultant vectors for different documents.

*6.2.4.3. Latent Semantic Analysis (LSA).*

LSA (Landauer et al., 2007) started with the same representation as the Vectorspace model - a term by document matrix with log entropy weighting. In order to reduce the contribution of noise to similarity ratings, however, the raw matrix is subjected to singular value decomposition (SVD). SVD decomposes the original matrix into a term by factor matrix, a diagonal matrix of singular values and a factor by document matrix. Typically, only a small number of factors (e.g., 300) are retained. To derive a vector representation of a novel document, term vectors are weighted, multiplied by the square root of the singular value vector and then added. As with the vector space model, the cosine is used to determine similarity.

#### *6.2.4.4. Sparse Nonnegative Matrix Factorization (SpNMF).*

Nonnegative Matrix Factorization (Xu et al., 2003) is a technique similar to LSA, which in this context creates a matrix factorization of the weighted term by document matrix. This factorization involves just two matrices - a term by factor matrix and a factor by term matrix - and is constrained to contain only nonnegative values. The nonnegative matrix factorization has been shown to create meaningful word representations using small document sets, in order to make it possible to apply it to large collections we implemented the sparse tensor method proposed by Shashua and Hazan (2005). As in LSA, log entropy weight term vectors were added to generate novel document vectors and the cosine was used as a measure of similarity.

### *6.2.5. Corpora*

#### *6.2.5.1. TASA.*

The Touchstone Applied Science Associates (TASA) corpus was constructed to represent the reading material that is covered by American students up to their first year of college. The 35471 documents contained in the TASA corpus range over nine content areas: language arts and literature, social science, science and math, fine arts, home economics and related fields, trade, service and technical fields, health, safety and related fields, business and related fields, popular fiction and nonfiction (Budiu, Royer, & Pirolli, 2007).

#### *6.2.5.2. Wikipedia.*

Wikipedia was used as a generic set of documents from which smaller targeted sub-spaces could be sampled and compiled. Wikipedia is maintained by the general public, and has become the largest and most frequently revised or updated encyclopedia in the world. Critics have questioned the accuracy of the articles contained in Wikipedia, however research conducted by Giles (2005) did not find significant differences in accuracy of science based articles in Wikipedia and similar articles contained in the Encyclopedia Britannica. In total, 2.8 million Wikipedia entries were collected and are current to March 2007. However, this

document number was reduced to 1.57 million after the removal of incomplete articles contained in the original corpus. The incomplete articles removed were identified if they contained the phrases like “help Wikipedia expanding” or “incomplete stub”.

To enable the creation of sub-space corpora, Lucene[5] (a high performance text search engine) was used to index each document in the Wikipedia corpus. Lucene allows the user to retrieve documents based on customized Boolean queries. These queries can include wild-card operators like ‘the star’ (*) to retrieve multiple results from word stems. Like the more well known search engine Google, the documents returned by Lucene are ordered by their relevance to a query.

The three search tasks given to the participants for each Web site were enumerated above in Section 2.3. Based on keywords selected from these tasks, the following Lucene queries (in italics) were constructed to retrieve the 1000 document sub-spaces from the Wikipedia corpus for each Web site.

Mission Australia:

*(“mission australia”) OR (“chief operating officer”) OR (“employment”) OR (“homeles*”)*

Green Corps:

*(“green corps”) OR (“national” AND “program” AND “manager”) OR (“environment*” AND “benefit*”) OR (“heritage” AND “benefit*”) OR (“partner” AND “agency”)*

White Lion:

*(“white lion”) OR (“company president”) OR (“organi?ation* president”) OR (“mentor*” AND “young” AND “people”) OR (“government department*” AND “support”)*

[5]PyLucene, is a Python extension that allows access to the Java version of Lucene: http://pylucene.osafoundation.org/

*6.2.5.3. Appending Web pages as Documents - The creation of TASA-WEB and WIKI-WEB corpora.*

When examining semantic models' ability to perform similarity estimates on paragraphs, Stone et al. (in press) found that including stimulus paragraphs into the semantic models' knowledge base (or corpus) provided an effective method of corpus improvement. Performance gains obtained in this way most likely result from context being given to novel stimulus terms that were not contained in the original corpus. To illustrate this point, the TASA corpus may not contain the phrase "Green Corps", but may contain the words "contribute" and "community". By appending documents (i.e., the text contained in Green Corps Web site) to the TASA corpus, the semantic model can more accurately make associations between these terms. Based on this Stone et al. (in press) finding, when creating corpora for each of the three Web sites in this study, the textual content of each of the Web pages viewed by participants on that Web site were appended to the TASA and Wikipedia corpora. For example, on the Mission Australia Web site, overall 57 unique Web pages were viewed by participants during the experiment. So, the textual content from these Web pages was used to construct a mini-corpus of 57 documents, which was then appended to both the TASA corpus and Wikipedia sub-spaces for Mission Australia corpora.[6] Highlighting this appended data, the naming conventions TASA-WEB and WIKI-WEB are used throughout this paper.

*6.2.6. Baseline models to estimate eye-position*

Three alternative models were designed as baselines to assess the success of the Semantic Fields models when predicting participants' eye-positions when they are engaged in goal-oriented search tasks on the three Web sites. These baseline models are the Flat, Non-Flat, and No-Model.

[6]Note: Green Corps and White Lion Web sites' textual data was not used when creating corpora for the Mission Australia data.

*6.2.6.1. Flat Model.*

The Flat model is the simplest model, it assumes that the eye-position has equal likelihood of being focused on all pixels contained on the Web page. Given the total number of eye-points (EP), and the total number of pixels that an eye-point ($p$) could be located in (1280 x 1024), for each page ($i$) that is viewed ($V$) by participants, the Flat model calculates the log-likelihood of the eye-points on any given Web page as $LL_{Webpage}$ (see Equation 3). The sum of these Web page log-likelihoods ($LL_{FLAT}$) calculates the fit of the Flat Model to all eye-points recorded for participants.

$$\begin{aligned} LL_{FLAT} &= \sum_{i \in V} LL_{Webpage_i} \\ LL_{Webpage} &= \sum_{p \in EP} \log\left(\frac{1}{1280 \times 1024}\right) \end{aligned} \tag{3}$$

*6.2.6.2. Non-Flat Model.*

The Non-Flat model is similar to the Flat model, with the exception that it gives more weight to the probability estimates for those eye-points found in textual elements.[7] The Non-Flat model is displayed in Equation 4, where for each Web page ($i$) that is viewed ($V$) by participants, $N$ is the number of pixels in text elements, $M$ is the number of pixels outside text elements, $A$ is the number of eye-points in text elements and $B$ is the number of eye-points outside text elements. Furthermore, $\hat{w}$ is the optimized probability of an eye-point being in a text element (see Equation 4). The maximized log-likelihood ($ML$) over all Web pages viewed by participants occurred at a MLE of $\hat{w} = 3.41$ for this sample. Therefore, participants were 3.41 times more likely to focus their eyes on the textual elements on a Web page than focusing on other areas. In accordance with this finding, the MLE has been used to calculate Non-Flat

[7] An example of text elements (including images that have a textual description) on a Web page are shaded in red in Figure 6.2.

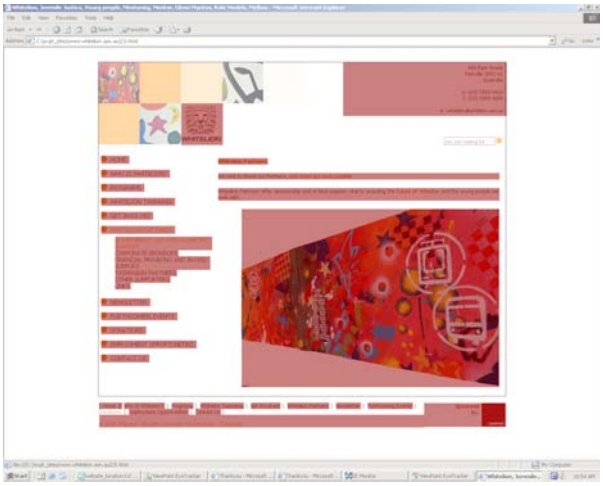


*Figure 6.2.* Textual Web page elements are highlighted in red, images that have "ALT" or descriptive text are included.

model log-likelihoods and assigned a greater weighting to eye-points recorded in these textual elements than the non-textual elements.

$$\begin{aligned} ML_{NONFLAT} = & \sum_{i \in V} ML_{Webpage_i} \\ ML_{Webpage} = & \; A \log\left(\frac{\hat{w}}{\hat{w}N + M}\right) \\ + & \; B \log\left(\frac{1}{\hat{w}N + M}\right) \end{aligned} \tag{4}$$

*6.2.6.3. No-Model.*

The No-Model condition has been created to test the theory that the Semantic Fields model is driven only by the structure of the Web page, and that the semantic models do not add to the Semantic Fields model's capacity to predict participants' eye-positions. It takes the same parameters as the Semantic Fields model, however the semantic model coefficient is kept

constant at one ($L_i = 1.0$) in the calculation of the Semantic Fields (see Equation 1).

*6.2.6.4. Calculating the log-likelihood for Semantic Fields and No-Model conditions.*

The log-likelihoods for the Semantic Fields models and No-Model are calculated in the same fashion. The Semantic Field value[8] ($SF_{(x,y)}$, see Equation 1) for each eye-point ($p$) is divided by the summed total of the Semantic Field values for all pixels that Web page ($SF_{Webpage}$). This process calculates the probability that a single eye point is viewed. Then, the log of these probabilities ($LL_{Webpage}$) is calculated and summed over all eye-points ($EP$) on a Web page viewed by a participant. This process of summing the log-likelihoods is repeated for each Web page ($i$) that was viewed ($V$) by participants to calculate the overall log-likelihood ($LL_{SF}$) for both the Semantic Fields and No-Model conditions (see Equation 5).

$$\begin{aligned} LL_{SF} &= \sum_{i \in V} LL_{Webpage_i} \\ LL_{Webpage} &= \sum_{p \in EP} \log\left(\frac{SF_p}{SF_{Webpage}}\right) \end{aligned} \tag{5}$$

## 6.3. Results

*6.3.1. Did the participants complete their tasks successfully?*

When designing the nine tasks, the researchers had identified specific destination Web pages that matched the information required by each task. Table 6.1 displays the percentage of times participants finished on the same Web page that was chosen by the experimenter. Overall, there was a 92.93 percent overlap between the expectations of the experimenter and the final landing pages chosen by participants. However, only half the participants completed the second task on the Green Corps Web site to the expectations of the researchers.

[8] No-Model holds the semantic model coefficient constant at 1.0, so the value it returns for a pixel can still be thought of as a Semantic Field value.

It seems likely that finding the environmental and heritage benefits contributed by Green Corps was more open to subjective judgments of completion by participants than were elicited by the other questions. On reviewing the data, it appears that simply finding the words 'environment' and 'heritage' on the same page may have prompted some participants to end their search. Although all tasks had been constructed with a specific end page in mind (see Figures 5-13 in Appendix F.1), on reflection the other eight tasks appear to have more specific goals. For example, finding the name of committee member or a specific file on the Web sites. While Green Corps Task 2 is a limitation of this study, overall all tasks the vast majority of participants found the target Web pages chosen by the experimenters.

Table 6.1: Percentage of overlap between the expected landing Web page chosen by the experimenter and those chosen by the 49 participants.

| | Task 1 | Task 2 | Task 3 | Average |
|---|---|---|---|---|
| Mission Australia | 100.00 | 95.92 | 95.92 | 97.28 |
| Green Corps | 95.65 | 53.06 | 97.92 | 82.21 |
| White Lion | 100.00 | 100.00 | 97.92 | 99.31 |
| Total | | | | 92.93 |

### *6.3.2. Were the participants paying attention?*

It is difficult to understand the psychophysiological connection between pupil dilation and cognitive load. However, researchers have been reporting this phenomenon for nearly a century now. The German psychiatrist Bumke, in his 1911 review of the relevant German literature, proposed that "in general every active intellectual process...produces pupil enlargement" (as cited by Hess, 1972, p. 492). A fairly consistent pattern of findings has been reported in the literature, with recordings of participants' pupil sizes increasing during more difficult tasks when compared to those recorded during easier tasks. This effect is often reported when participants are engaged in mathematical problem solving of varying

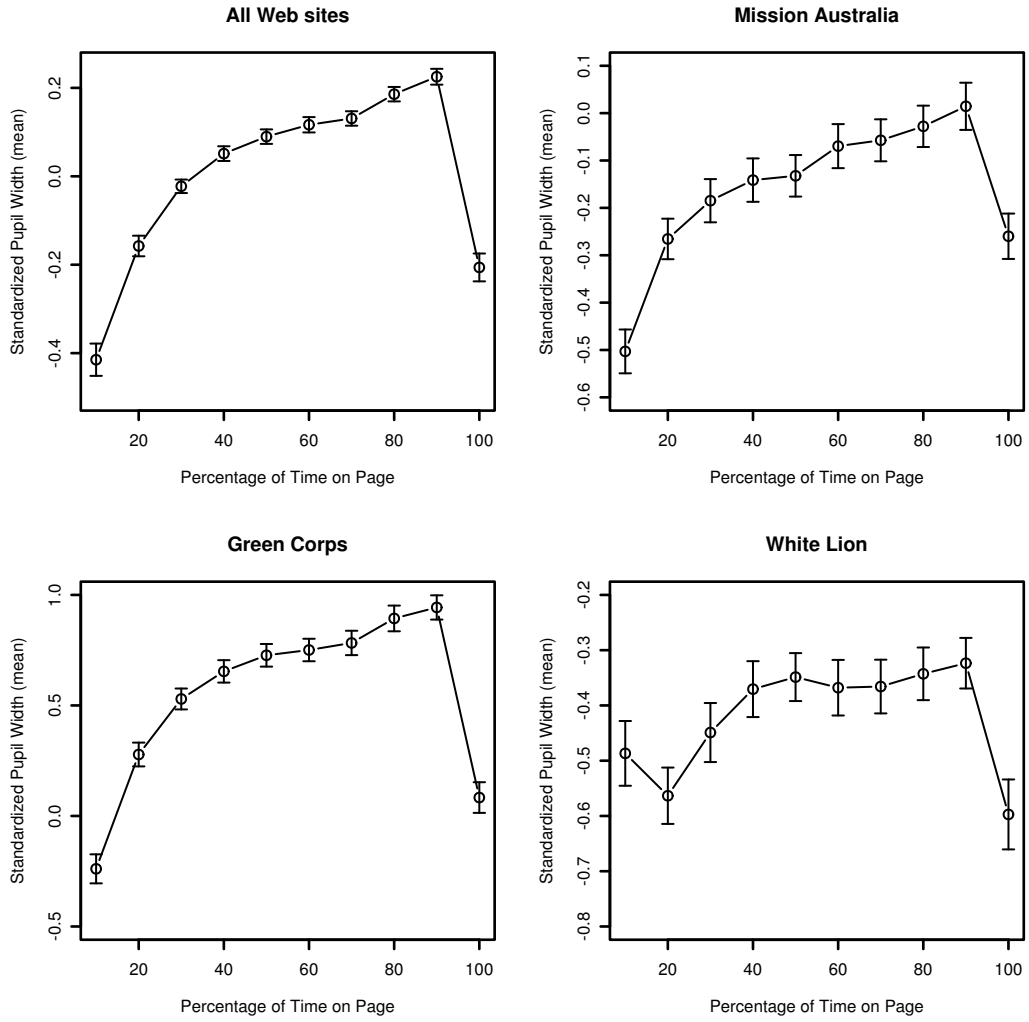



*Figure 6.3*. Standardized pupil width during participants' fixations while they were performing goal-oriented Web page navigation. Time spent searching each page is delineated into deciles.

difficulty (Stone et al., 2004; Steinhauer et al., 2000; Ahern & Beatty, 1979; Boersma et al., 1970; Schaeffer et al., 1968; Hess & Polt, 1964). In a preparatory study, Stone et al. (2004) found that under similar experimental conditions to those used in this study (eye-tracker, CRT monitor, room and lighting), participants' pupil sizes were larger whilst engaged in more difficult mathematical subtraction tasks when compared to their pupil size during easier adding tasks. Moreover, this pattern was consistently found when participants undertaking these tasks

were directed to focus on twenty-five evenly spaced targets, that were positioned inside the cells of a five-by-five grid on the CRT monitor. This study indicated that greater cognitive load could be detected using the pupil width measure across a wide range of focal points on the CRT monitor.

Figure 6.3 displays participants' pupil width data both overall and on each Web site in relation to the time spent searching each Web page. Using eye data collected on all three Web sites, pupil width Z-scores were calculated individually for each participant. This standardization of participants' pupil widths was used to control for individual differences in pupil size. Furthermore, participants' reading speed, cognitive processing speeds and time spent on any given page varies greatly, therefore to control for these variations, time spent on the page has been divided in deciles, as opposed to presenting it in seconds which would have confounded these individual differences in participants' performance.

It is interesting to note that in all nearly cases participants' pupil width has increased over the time spent on each Web page. The only deviation to this trend is shown on the White Lion Web site, where pupil size decreases in the first 20 percent on time spent on the Web pages. Greater luminance causes the participants' pupils to constrict. Therefore, it is likely that the decrease in participants' pupil size at the beginning of each White Lion page can be attributed to the brighter pages on the White Lion Web site when compared to the lower luminance calibration display.[9] This conclusion is also supported by the lower z-scores recorded on the White Lion Web site when compared to either the Green Corps (darkest) or Mission Australia (mid range) Web sites.

Assuming pupil dilation is a measure of the participants' cognitive load, then larger standardized values will reflect greater cognitive processing by the participant. In Figure 6.3, it is clearly shown that as participants' spend more time on each Web page, the width of the pupil

[9] As mentioned previously, the calibration display (set on a gray background) was presented prior to each Web page viewed by participants.

increases. This suggests that as participants' spent more time on each page, their cognitive load also increased. It is likely that these cognitive load increases are related to the decision processes undertaken by the participants as they assess an appropriate navigation path to complete their search goal (e.g., what information is important on the Web page, and whether to click on to a more relevant page).

After the ninth decile pupil width decreases in all cases. This probably reflects the time period after which participants' have either found the target or have clicked on a link[10] to pursue and are waiting for the browser to load the next page. In both of these cases the next page participants are taken to is the calibration page. Moreover, in both scenarios one would expect a reduction in participants' cognitive load during this time when compared to that produced by participants during time spent reading and assessing page content for information relevant to search goals.

### *6.3.3. Ten models compared using the Bayesian Information Criterion*

The log-likelihoods of ten models (see Equations 3-5) constructed to predict participants eye movements are compared in this section using the Bayesian Information Criterion (BIC, Schwarz, 1978) for assessing model fit. The ten models include seven Semantic Fields (SF) models. One of these SF models, word overlap, does not use a knowledge base. Evaluations are only made on the co-occurrences between words in Web page element text and the textual description of user goals. The other six SF models are more complex and use backgrounding documents, half of these SF models used the TASA-WEB corpus while the other half used WIKI-WEB corpora as a knowledge base. Furthermore, three semantic models (LSA, SpNMF, and Vectorspace) were used to compare goal text to element text in these six SF models. So, for the six more complex SF models, there is a two (corpus) by three (semantic model) experimental design. The other three models compared here are the Flat, Non-Flat, and Non-

[10] It is not possible to verify this in our data, as a time stamp was not generated for the mouse click action.

Model conditions. While no parameters are set in the log-likelihood calculations for most of these models, the maximized log-likelihood equation for the Non-Flat model has one parameter ($\hat{w}$). The BIC is an appropriate method for comparing the fit of these models to the eye-data, because it adjusts for the number of parameters going into the model. Moreover, higher BIC scores indicate a better fitting model to the data.

Table 6.2: Comparison of Bayesian Information Criteria (BIC) statistics calculated from log-likelihoods generated for all ten models.

| Corpora | Model | BIC | Difference |
|---|---|---|---|
| WIKI-WEB | Vectorspace | -10983963 | 0 |
| WIKI-WEB | SpNMF | -10989551 | -5588 |
| WIKI-WEB | LSA | -10990633 | -6670 |
| TASA-WEB | Vectorspace | -10993390 | -9427 |
| TASA-WEB | SpNMF | -11001041 | -17078 |
| TASA-WEB | LSA | -11001286 | -17323 |
| - | No-Model | -11005111 | -21148 |
| - | Overlap | -11094311 | -110348 |
| - | Non-Flat | -11287790 | -303827 |
| - | Flat | -11417562 | -443599 |

The results displayed in Table 6.2 are presented in descending order of their BIC scores. There were three interesting trends displayed in the results. Firstly, BICs were higher for the six corpus-based Semantic Fields models than the other models, therefore the Semantic Fields models provided a better fit for the eye tracking data than the other baseline models. Moreover, as would be expected, the simple Flat model performed the worst, followed by the Non-Flat model, and then the No-Model; the latter two baseline models were expected to perform better as they contain information about the structure of the Web page display. Secondly, corpus choice appeared to affect SF model performance. The SF models using the WIKI-WEB corpora outperformed the TASA-WEB corpora in all instances. Thirdly, the semantic models that were used in the SF models produced a main effect. Using either corpus as a knowledge

base, Vectorspace was consistently the best performing model, followed by SpNMF and then LSA.

It is unsurprising that the word overlap model performed badly. Unlike the corpus-based semantic models, word overlap will only find similarity when there is an overlap of words between textual elements on a Web page and the textual description of a task. Alternatively, the models such as Vectorspace, LSA and SpNMF have a greater pool of words and their associates contained in the corpus to draw on when making these similarity estimates. Finding even a small degree of similarity, based on an associated word (i.e., not the exact matching of words required by the word overlap model), will generate SF model 'heat' from its source location within the structure of a Web page. That is, if there is no association found between the task and the text in a Web page element, then the presence of that element (even if it is a menu item), cannot inform the SF model. Furthermore, the importance of including structural information into the SF model is highlighted by the better performance of the No-Model condition (which only uses Web page structure as opposed to semantics) in comparison to the word overlap model.

### *6.3.4. How well does the VEC-SF model using the WIKI-WEB corpus predict the eye data?*

During complex tasks that require reading, such as goal-oriented Web page search, the location of participants' eye fixations are generally an indicator of goal-specific attentional processes (p. 375 Rayner, 1998). To accomplish their task, participants engaged in visual search must seek out locations of goal-relevant information on the display. If the Semantic Fields model is able to capture some of the variability in participants' eye movements, then one would expect that Semantic Field values located at participants' eye focal points will increase as participants' spend more time viewing a page.

As reported above, the Semantic Fields model that used Vectorspace with the WIKI-WEB corpus provided the best performance when estimating the location of participants' eye-

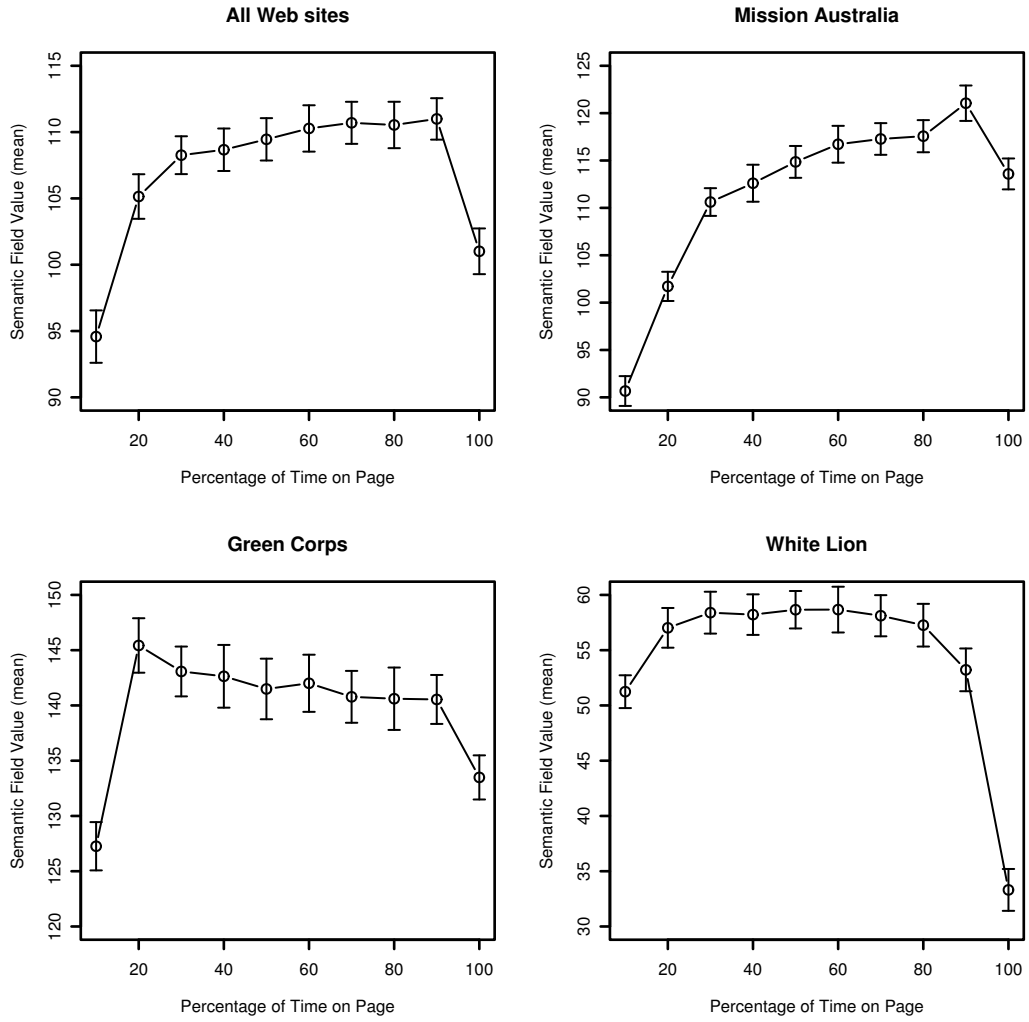



*Figure 6.4*. Semantic Field values (Vectorspace with the WEB-WIKI corpus) calculated for participant eye-points during goal-oriented Web page navigation. Time spent searching each page is delineated into deciles.

points during Web navigation. To provide the reader with a visual illustration of the Semantic Fields generated using both Vectorspace and the WIKI-WEB corpus, Semantic Field maps for the target pages for all nine tasks are displayed in Figures 5-13 in Appendix F.1. Figure 6.4 displays the mean SF values using Vectorspace with the WIKI-WEB corpus for each page viewed by participants. Following the same line of reasoning that was outlined in Section

3.2, time spent on page has been divided into deciles to avoid possible confounds produced by individual differences in participants' task performance. On all three Web sites, it appears that after an initial orienting phase, participants' eye-movements move towards areas of greater SF values. As mentioned in the pupil section, it is possible that the sharp drop off in the last 10% of time spent on the page, captures eye movement between the click of a link and the browser moving onto the next page.

## 6.4. Discussion

In this paper we have evaluated the Semantic Fields models' ability to estimate the location of 49 participants' eye movements during goal-oriented search tasks on three Web site. It was found that all corpus-based Semantic Fields models performed better than other models that depended solely on display characteristics. Particularly encouraging was finding that the Semantic Fields models outperformed No-Model condition. The No-Model condition is essentially the Semantic Fields model with textual similarity estimates held constant at one. Incorporating the same decay function as the Semantic Fields model, the No-Model condition produces heat solely based on the location of Web page elements. That the Semantic Fields model outperforms the No-Model condition indicates that the semantic component of the Semantic Fields model is providing more to the fit of this model to the eye-tracking data than can be produced by the display component alone. While the Semantic Fields models provided the best fit to the human eye-tracking data in this study, there were some interesting performance differences between each of the corpus-based Semantic Fields models. These performance differences were introduced by the manipulation of both corpora and semantic models used by the Semantic Fields model.

LSA has been the focus of much of the statistical lexical semantics research in recent years. That LSA has successfully been used to grade essays (Foltz et al., 1999) is a testament to the overall usefulness of this model. It was therefore surprising that a much simpler model

like Vectorspace, would consistently out-perform more complex models such as LSA and SpNMF when generating similarity comparison of text in this study. It is also interesting to note that Vectorspace is the first step in the calculation of both LSA and SpNMF, which begs the question as to whether the extra complexity introduced by these latter models when employing dimensionality reduction is of benefit when performing textual comparisons of user goals and Web page content. Furthermore, the simplicity of Vectorspace's calculation allows for quick and efficient construction of "on-the-fly" semantic knowledge spaces that could be incorporated into more applied models of semantic salience on Web pages.

Drawing targeted corpora for the semantic models from the larger corpus of Wikipedia provided better knowledge bases for the semantic models than a more traditional corpus like TASA. The TASA corpus has been hand-picked to broadly represent the expected general knowledge of a first-year American college student (Dennis, 2007). However, the findings in this study indicate that for some semantic models (Vectorspace, LSA and SpNMF), semi-automated corpora generation using Wikipedia provides a better base to compare the similarity of textual information. That said, the generation of Wikipedia sub-spaces in this research was based on very simple Boolean queries. Greater focus on the formulation of these Lucene queries may increase the performance of semantic models when calculating text similarity and thereby conceivably produce better estimates of eye-tracking data by the Semantic Fields model.

The pupil width measure indicated that participants' cognitive load increased as they spent more time on each Web page. Overall, most participants ended their goal-oriented search on the same pages that were initially identified by the experimenters when constructing the search tasks. Taken together, these findings support the notion that participants were actively searching out their target goals during this experiment. Given that participants were predominantly "on-task" during this experiment, it was also pleasing to find that the corpus-based Semantic Fields models were able to outperform the other models under these

circumstances.

## 6.5. Conclusions

Both Web page semantics and display characteristics determine the success with which a user will be able to find information on a Web page. The Semantic Fields model incorporates both of these characteristics, and was found to provide better estimates of participants' eye-movements during goal-oriented search than could be generated by solely display-based models. Choices of both the semantic model and knowledge base affected the performance of the semantic component that is used by the Semantic Fields model. Contrary to expectations, a relatively simple semantic model, Vectorspace, outperformed more complex semantic models that employ dimensionality reduction. Also, better approximations to the knowledge required to successfully estimate textual similarity were produced by targeted corpora drawn from Wikipedia when compared to those found using the more generic TASA corpus. Overall, the Semantic Fields model that used both Vectorspace and a targeted corpus drawn from Wikipedia, was found to be the best performing model when estimating participants' eye movements during goal-oriented search tasks in this study.

# Chapter 7. General Conclusion

This thesis has described the development and assessment of the Semantic Fields model of visual salience during goal-oriented Web site search tasks. Four papers were presented in Chapters 3-6 that document the process of developing and assessing this model. In Paper 1 (see Chapter 3), pupil dilation was validated as a measure of cognitive load for use in later studies. Paper 2 (see Chapter 4) reported on the first attempt to use the Semantic Fields model to estimate data collected from participants who were engaged in goal-oriented search on three Web sites. In Paper 3 (see Chapter 5), four studies were presented in which the semantic component of the Semantic Fields model was refined. Finally, in Paper 4 (see Chapter 6), performance at estimating visual salience was compared between seven versions of the Semantic Fields model and three solely display-based models on the human goal-oriented Web site search dataset.

The Semantic Fields model provided the most useful estimates goal-oriented of visual salience on Web pages. As participants' sort out Web page regions that more closely resembled their goals, it was found that they focused on areas of greater visual salience estimated by the Semantic Fields model (see Section 6.3.4). Furthermore, the Semantic Fields model outperformed all of the solely display-based models of visual salience used in this research (see Sections 4.3 and 6.3.3). While these are important findings to the validation of the Semantic Fields model, there may be limitations to their generalizability. The reader should bear in mind that there were only three Web sites examined by participants. Also, while there were many pages viewed by each participant, each Web site and its pages had a corporate layout. Therefore, it should be noted that these findings may not generalize to other Web sites that incorporate different display formats.

It was particularly encouraging to find that both Web page semantics and display characteristics determined the success with which the Semantic Field models estimated goal-

oriented visual salience. Initially, it was feared that the semantic component of the Semantic Fields model was not informing the model's estimates by much more than could be generated using only the display component of the model. This motivated a thorough assessment of the factors influencing the semantic component's ability to make similarity estimates between texts (see Chapter 5). The findings of this assessment led to the construction of an improved version of the semantic component of Semantic Fields model. Allaying previous concerns, this improved Semantic Fields model provided better estimates of visual salience than could be provided by the display component of the Semantic Fields model alone (see Section 6.3.4).

The examination of the semantic component of the Semantic Fields model produced two consistent results. Both choice of semantic model and corpus influenced the semantic component of the Semantic Fields model's ability to perform similarity estimates between goal and Web page element texts. In the assessment of the semantic component, Salton et al.'s (1975) vector space model provided the best estimate of textual similarity (0.67) when compared to the judgments made by human raters (see Chapter 5). This finding was further supported in the subsequent visual salience modeling using the Semantic Fields model presented in Paper 4. To this end, the best estimates of visual salience on Web pages found in this research were provided by the Semantic Fields model which used the vector space model to inform its semantic component.

It was interesting that the vector space model performed so well in both areas of this research. Particularly as the vector space model was the simplest of the corpus-based semantic models. While models such as LSA and SpNMF also performed well, they did not perform as well as the vector space model and are both more computationally expensive and difficult to implement. The vector space model is employed in the initial stages of LSA, which is then followed by dimensionality reduction of the vector space model's term vectors. This dimensionality reduction merges information across term vectors to retrieve higher order constructs or factors in that document set. Therefore, it is notable that the dimensionality

reduction conducted by these models has not generally improved their performance when estimating textual similarity in this research program.

One possible explanation for this outcome is suggested by Louwerse (in press), who has proposed that support for language comprehension and production is based outside the user, and is instead situated in language itself. In arguing his claim, Louwerse provides examples of how first-order co-occurrences of terms can produce similar results to LSA on tasks of categorization. Similarly, it could be argued that to some extent the good performance of the overlap model in Paper 3 and the vector space model in both Papers 3 and 4 support Louwerse's argument.

Another explanation for the success of the overlap and Vectorspace models in both Papers 3 and 4 may be found in the framework of the tasks set to participants. More specifically, skimming as a reading strategy offers a potential mechanism that may be better captured by these simpler semantic models than the more complex models that incorporate dimensionality reduction. These simpler models rely heavily on the co-occurrence of words to make their similarity estimate, rather than attempting to address deeper underlying themes or concepts that may manifest from a knowledge base.

For example, in Paper 3 the side-by-side positioning of these paragraph pairs in the experimental procedures may have encouraged skimming and keyword-matching (or discrimination) between these paragraphs by the participants. Moreover, it is not hard to imagine that a time-starved or disinterested participant may resort too skimming as a strategy to expediate the experimental task at hand. In the discussion of Paper 3, it was noted that Masson (1982) found visually unique features of text such as place names may increase efficiency of skimming as a reading strategy. Given that proper nouns were present in all of the paragraphs presented to participants in this reseach, commonalities between participants' similarity estimates (and also those of the overlap and Vectorspace models) may also be influenced or encouraged by the co-occurrence of these unique entities between stimulus

paragraphs.

Similarly, several researchers have reported that Web users skim Web pages for target information (Blackmon, Kitajima, & Polson, 2003; Spool, Schroeder, Scanlon, & Snyder, 1998; Nielsen, 1997). Therefore, it is possible that the vector space model's better estimates of similarity between the search goal and the text displayed on a Web page are a reflection of the Web users' skim reading or scanning for information. This proposition is supported by the greater variety of goal destination pages recorded for Task 2 on the Green Corps Web site. As mentioned in Paper 4, in retrospect Task 2 was more open to subjective interpretation by the participants, and many participants seemed to conclude their searches on Web pages that contain only keywords from the task goal. If this is the case, then a deeper reading of the Web pages by participants may have been associated with better performance by the Semantic Fields models that used LSA or SpNMF. However, given that Web page users may skim for relevant information on Web page, then the vector space model may correctly provide the Semantic Fields model with the best fit to the human data.

The knowledge base used by the semantic component also influence the performance of the Semantic Fields model. Web site specific, goal-targeted sub-corpora drawn from Wikipedia provided the best knowledge base for the Semantic Fields models (see Section 6.3.3). Moreover, this finding occurred across all of the corpus-based semantic models that informed the semantic component of the Semantic Fields model. Given the vast array of content that can be found on the World Wide Web, it is almost inconceivable that a generic knowledge base such as TASA could adequately inform a model of visual salience that requires estimation of textual similarity. However, corpora that are hand-picked for a particular knowledge domain can be both difficult and time consuming to construct. Therefore, it was encouraging to find that semi-automated generation of sub-corpora from Wikipedia provided a relative good match to the semantic knowledge required to perform the paragraph comparisons in Paper 3 and the estimates of goal-oriented textual similarity used by the Semantic Fields

models' in Paper 4.

Besides the Wikipedia corpus covering a larger range of topics than corpora like the TASA, the Wikipedia sub-corpora tended to contain more unique words per document (see Table E.4.1) than other corpora used in this research. Also, the method of appending stimulus documents to the Wikipedia sub-corpora may increase this vocabulary and give greater context to stimulus specific words and phrases. This greater context and vocabulary for each concept is likely to have contributed to the performance of the semantic models when performing similarity estimates between texts.

It was also interesting to note the differences in motivation or cognitive load that were assessed by the pupil measure in this research. Participants cognitive load was seen to decrease in both mathematical tasks (see Figure 2.1). However, when performing the goal-oriented search on Web sites the participants' cognitive load increased with the time they spent on each Web page (see Figure 6.3). At one level this is encouraging because the participants appeared to be 'on task' when performing the goal-oriented Web site search tasks. At another level, these results offered an accurate reflection of the tasks given to the participants. It is easy to imagine why participants' might become bored or lose track of their count performing a repetitive mathematical task for two minutes. Moreover, this type of task is not generally performed during people's day to day lives. On the other hand, searching a Web site for information is a task most students and academics might perform on a daily basis. Therefore, participants' greater engagement or motivation on the Web site search tasks, may reflect the greater ecological validity of the Web-based experiment performed in this research. To this end, the participants were provided with a familiar interface (Internet Explorer), and asked to search on professionally developed 'active' Web sites for plausible search goals. Under these conditions, and with motivated participants, it was satisfying that the Semantic Fields model provided reasonable estimates of the goal-related visual salience.

## *7.1. Final Statement*

Both Web page semantics and display characteristics determine the success with which a user will be able to find information on a Web page. The Semantic Fields model incorporates and is informed by both of these characteristics. Based on this research program, it is concluded that the Semantic Fields model provided useful estimates of visual salience during participants' goal-oriented search of Web sites.

# References

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing varying with scholastic test scores. *Science*, *205* (4412), 1289–1292.

Appenzeller, O., & Oribe, E. (1997). *The autonomic nervous system: An introduction to the basic and clinical concepts*. Amsterdam: Elsevier.

Arlington Research, Inc. (2002). *ViewPoint EyeTracker ™PC-60 Software Users Guide.*

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91* (2), 276–292.

Beatty, J., & Wagoner, B. L. (1978). The pupillary response as an indicator of arousal and cognition. *Science*, *4334* (199), 1216–1218.

Bell, N., & Garland, M. (2008). *Efficient sparse matrix-vector multiplication on CUDA* (NVIDIA Technical Report No. NVR-2008-004). NVIDIA Corporation. Available at: http://www.nvidia.com/object/nvidia_research_pub_001.html Accessed April 10, 2009.

Berry, M. W., & Browne, M. (2005). Email surveillance using non-negative matrix factorization. *Computational & Mathematical Organization Theory*, *11* (3), 249–264.

Blackmon, M. H., Kitajima, M., & Polson, P. G. (2003). Repairing usability problems identified by the cognitive walkthrough for the web. In G. Cockton & P. Korhonen (Eds.), *Chi '03: Proceedings of the sigchi conference on human factors in computing systems* (pp. 497–504). New York, NY, USA: ACM.

Blackmon, M. H., Kitajima, M., & Polson, P. G. (2005). Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. In G. van der Veer & C. Gale (Eds.), *CHI '05: Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 31–40). New York, NY, USA: ACM.

Blackmon, M. H., Mandalia, D. R., Polson, P. G., & Kitajima, M. (2007). Automating usability evaluation: Cognitive Walkthrough for the Web puts LSA to work on real-world HCI design problems. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 345–375). Mahwah, NJ: Lawrence Erlbaum Associates.

Blackmon, M. H., Polson, P. G., Kitajima, M., & Lewis, C. (2002). Cognitive walkthrough for the web. In D. Wixon & L. Terveen (Eds.), *CHI '02: Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 463–470). New York, NY, USA: ACM.

Blei, D., Ng, A. Y., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3* (4-5), 993–1022.

Boersma, F., Wilton, K., Barham, R., & Muir, W. (1970). Effects of arithmetic problem difficulty on pupillary dilation in normals and educable retardates. *Journal of Experimental Child Psychology*, *9* (2), 142–155.

Bradshaw, J. L. (1968). Pupil size and problem solving. *The Quarterly Journal of Experimental Psychology*, *20* (2), 116–122.

Brumby, D. P., & Howes, A. (2003). Interdependence and past experience in menu choice assessment. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th annual conference of the Cognitive Science Society* (p. 1320). Mahwah, NJ: Lawrence Erlbaum Associates.

Brumby, D. P., & Howes, A. (2004). Good enough but I'll just check: Web-page search as attentional refocusing. In M. C. Lovett, C. D. Schunn, C. Lebiere, & P. Munro (Eds.), *Proceedings of the sixth international conference on cognitive modeling* (pp. 46–51). Mahwah, NJ: Lawrence Erlbaum Associates.

Budiu, R., Royer, C., & Pirolli, P. (2007). Modeling information scent: a comparison of LSA, PMI and GLSA similarity measures on common tests and corpora. In *Proceedings of the 8th annual conference of the Recherche d'Information Assistée par Ordinateur (RIAO).* Pittsburgh, PA: Centre des Hautes Études Internationales d'Informatique Documentaire.

Bullinaria, J. A., & Levy, J. P. (2006). Extracting semantic representations from word co-occurrence statistics: a computational study. *Proceedings of the National Academy of Sciences*, *39* (3), 510–526.

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, *12* (2-3), 177–210.

Cai, Z., McNamara, D. S., Louwerse, M. M., Hu, X., Rowe, M., & Graesser, A. C. (2004). NLS: A Non-Latent Similarity Algorithm. In D. G. Forbus & T. Regier (Eds.), *Proceedings of the 26th annual conference of the Cognitive Science Society* (pp. 180–185). Mahwah, NJ: Lawrence Erlbaum Associates.

Carroll, J. S., & Johnson, E. J. (1990). *Decision research: A field guide*. Newbury Park: Sage Publications.

Chi, E. H., Pirolli, P., Chen, K., & Pitkow, J. (2001). Using information scent to model user information needs and actions and the web. In M. Beaudouin-Lafon & R. J. K. Jacob (Eds.), *CHI '01: Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 490–497). New York, NY: ACM.

Chi, E. H., Rosien, A., Supattanasiri, G., Williams, A., Royer, C., Chow, C., et al. (2003). The bloodhound project: automating discovery of web usability issues using the InfoScent$\pi$ simulator. In G. Cockton & P. Korhonen (Eds.), *CHI '03: Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 505–512). New York, NY, USA: ACM.

Choueka, Y., & Lusignan, S. (1985). Disambiguation by short contexts. *Computers and the Humanities*, *19* (3), 147–157.

Christie, B. (1985). *Human factors of the user-system interface: A report on an ESPRIT preparatory study*. New York, NY: Elsevier.

Cox, A., & Young, R. M. (2004). A rational model of the effect of information scent on the exploration of menus. In M. C. Lovett, C. D. Schunn, C. Lebiere, & P. Munro (Eds.), *Proceedings of the sixth*

*international conference on cognitive modeling* (pp. 82–87). Mahwah, NJ: Lawrence Erlbaum Associates.

Dennis, S. (2005). A memory-based theory of verbal cognition. *Cognitive Science*, *29* (2), 145–193.

Dennis, S. (2007). How to use the LSA web site. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 57–70). Mahwah, NJ: Lawrence Erlbaum Associates.

Dennis, S., Bruza, P., & McArthur, R. (2002). Web searching: A process-oriented experimental study of three interactive search paradigms. *Journal of the American Society for Information Science and Technology*, *53* (3), 120–133.

Faraday, P. (2000). Visually critiquing web pages. In *Proceedings of the 6th conference on Human Factors and the Web. Available at: http://www.tri.sbc.com/hfweb/faraday/faraday.htm Accessed October 20, 2006.*

Faraday, P. (2001). Attending to web pages. In *Chi '01: Chi '01 extended abstracts on Human Factors in computing systems* (pp. 159–160). New York, NY, USA: ACM.

Fitts, P. M., Jones, R. E., & Milton, J. L. (1950). Eye movement of aircraft pilots during instrument-landing approaches. *Aeronautical Engineering Review*, *9* (2), 24–29.

Foltz, P. W. (2007). Discourse coherence and LSA. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 167–184). Mahwah, NJ: Lawrence Erlbaum Associates.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, *1* (2). Available at: http://imej.wfu.edu/articles/1999/2/04/index.asp Accessed April 2, 2008.

Fu, W.-T., & Pirolli, P. (2007). SNIF-ACT: a cognitive model of user navigation on the World Wide Web. *Human-Computer Interaction*, *22* (4), 355–412.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In M. M. Veloso (Ed.), *Proceedings of the 20th international joint conference on artificial intelligence* (pp. 1606–1611). Menlo Park, CA: AAAI Press.

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 75–96). Oxford, UK: University Press.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *438* (7070), 900–901.

Granka, L., Hembrooke, H., & Gay, G. (2006). Location location location: Viewing patterns on www pages. In K.-J. Räihä & A. T. Duchowski (Eds.), *ETRA 2006: Proceedings of the 2006 symposium on eye tracking research & applications* (p. 43). New York, NY: ACM.

Granka, L., Hembrooke, H., Gay, G., & Feusner, M. (2004). *Correlates of visual salience and disconnect.* Unpublished research report, Cornell University Human-Computer Interaction Lab. Available at: http://www.hci.cornell.edu/projects/eye_tracking.htm Accessed April 29, 2007.

Grier, R. A. (2004). *Visual attention and web design*. Unpublished doctoral dissertation, University of Cincinnati, Department of Psychology. Available at: http://etd.ohiolink.edu/view.cgi?ucin1092767744 Accessed October 20, 2006.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th annual conference of the Cognitive Science Society* (pp. 381–386). Mahwah, NJ: Lawrence Erlbaum Associates.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101 (Suppl. 1)*, 5228–5235.

Griffiths, T. L., Tenenbaum, J. B., & Steyvers, M. (2007). Topics in semantic representation. *Psychological Review*, *114* (2), 211–244.

Habuchi, Y., Kitajima, M., & Takeuchi, H. (2008). Comparison of eye movements in searching for easy-to-find and hard-to-find information in a hierarchically organized information structure. In K.-J. Räihä & A. T. Duchowski (Eds.), *ETRA 2008: Proceedings of the 2008 symposium on eye tracking research & applications* (pp. 131–134). New York, NY: ACM.

Habuchi, Y., Takeuchi, H., & Kitajima, M. (2006). The influence of web browsing experience on web-viewing behavior. In K.-J. Räihä & A. T. Duchowski (Eds.), *ETRA 2006: Proceedings of the 2006 symposium on eye tracking research & applications* (pp. 47–47). New York, NY: ACM.

Hess, E. H. (1972). Pupillometrics: A method of studying mental, emotional, and sensory processes. In N. S. Greenfield & R. A. Sternbach (Eds.), *Handbook of psychophysiology* (pp. 491–531). New York: Holt, Rinehart and Winston.

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, *132* (3423), 349–350.

Hess, E. H., & Polt, J. M.(1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143* (3611), 1190–1192.

Hyönä, J., Tommola, J., & Alaja, A.-M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology Section A*, *48* (3), 598–612.

Jorna, P. G. A. M. (1997). Human machine interfaces for atm: objective and subjective measurements on human interactions with future flight deck and air traffic control systems. Available at: http://atm-seminer-97.eurocontrol.fr/jorna.htm Accessed October 2, 2004.

Josephson, S., & Holmes, M. E. (2002). Attention to repeated images on the world-wide web: Another look at scanpath theory. *Behavior Research Methods, Instruments, & Computers*, *34* (4), 539–548.

Kahneman, D.(1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154* (3756), 1583–1585.

Kaur, I., & Hornof, A. J. (2005). A comparison of LSA, wordNet and PMI-IR for predicting user click behavior. In G. van der Veer & C. Gale (Eds.), *CHI '05: Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 51–60). New York, NY, USA: ACM.

Kintsch, W.(2001). Predication. *Cognitive Science*, *25* (2), 173–202.

Kireyev, K.(2008). Beyond words: Semantic representation of text in distributional models of language. In M. Baroni, S. Evert, & A. Lenci (Eds.), *Proceedings of the ESSLLI workshop on distributional lexical semantics: Bridging the gap between semantic theory and computational simulations* (pp. 25–33). Hamburg, Germany: ESSLLI.

Kitajima, M., Blackmon, M. H., & Polson, P. G. (2000). A comprehension-based model of Web navigation and its applications to Web usability analysis. In S. McDonald, Y. Waern, & G. Cockton (Eds.), *People and computers XIV: Usability or else! Proceedings of HCI 2000* (pp. 357–373). New York: Springer.

Kitajima, M., Blackmon, M. H., & Polson, P. G. (2005). Cognitive architecture for website design and usability evaluation: comprehension and information scent in performing by exploration. In *HCI International 2005, vol. 4, Theories, Models and Processes in HCI.* CD-ROM (ISBN 0-8058-5807-5). Mahwah, NJ: Lawrence Erlbaum Associates.

Kommerell, G., Schmitt, C., Kromeier, M., & Bach, M. (2003). Ocular prevalence versus ocular dominance. *Vision Research*, *43* (12), 1397–1403.

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin and Review*, *12* (4), 703–710.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104* (2), 211–240.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lee, M. D., & Corlett, E. Y. (2003). Sequential sampling models of human text classification. *Cognitive Science*, *27* (2), 159–193.

Lee, M. D., Pincombe, B. M., & Welsh, M. B. (2005). An empirical evaluation on models of text document similarity. In B. G. Bara, L. W. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the Cognitive Science Society* (pp. 1254–1259). Mahwah, NJ: Lawrence Erlbaum Associates.

Lin, Y., Zhang, W. J., & Watson, L. G. (2003). Using eye movement parameters for evaluating human-machine interface frameworks under normal control operation and fault detection situations. *International Journal of Human-Computer Studies*, *59* (6), 837–873.

Ling, J., & Van Schaik, P. (2002). The effect of text and background color on visual searches of web pages. *Displays*, *23* (5), 223–230.

Ling, J., & Van Schaik, P. (2004). The effect of link format and screen location on visual search of web pages. *Ergonomics*, *47* (8), 907–921.

Louwerse, M. M. (in press). Symbol interdependency symbolic and embodied cognition. *Topics in Cognitive Science*.

Martin, D. I., & Berry, M. W. (2007). Mathematical foundations behind Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 35–55). Mahwah, NJ: Lawrence Erlbaum Associates.

Martin, M. J., & Foltz, P. W. (2004). Automated team discourse annotation and performance prediction using LSA. In S. T. Dumais, D. Marcu, & S. Roukos (Eds.), *HLT-NAACL 2004: Short papers* (pp. 97–100). Boston, Massachusetts, USA: Association for Computational Linguistics.

Masson, M. E. J. (1982). Cognitive processes in skimming stories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8* (5), 400–417.

McCarthy, J., Sasse, M. A., & Rigelsberger, J. (2003). Could I have the menu please? An eye tracking study of design conventions. In E. O'Neill, P. Palanque, & P. Johnson (Eds.), *People and computers XVII: Designing for society. proceedings of HCI 2003* (pp. 401–414). London, UK: Springer-Verlag.

McNamara, D. S., Cai, Z., & Louwerse, M. M. (2007). Optimizing LSA measures of cohesion. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 379–400). Mahwah, NJ: Lawrence Erlbaum Associates.

Miller, C. S., & Remington, R. W. (2004). Modeling information navigation: Implications for information architecture. *Human-Computer Interaction*, *19* (3), 225–271.

Miller, G. A. (1995). WordNet: A lexical database for english. *Communications of the ACM*, *38* (11), 39–41.

Moed, H. F., Glänzel, W., & Schmoch, U. (2004). *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems*. Secaucus, NJ, USA: Springer–Verlag New York.

Nakayama, M., Takahashi, K., & Shimizu, Y. (2002). The act of task difficulty and eye-movement frequency for the 'oculo-motor indices'. In A. T. Duchowski, R. Vertegaal, & J. W. Senders (Eds.), *ETRA '02: Proceedings of the 2002 symposium on Eye Tracking Research & Applications* (pp. 37–42). New York, NY: ACM.

Namatame, M., & Kitajima, M. (2008). Suitable representations of hyperlinks for deaf persons: an eye-tracking study. In S. Harper & A. Barreto (Eds.), *Assets '08: Proceedings of the 10th*

*international ACM SIGACCESS conference on computers and accessibility* (pp. 247–248). New York, NY, USA: ACM.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The university of South Florida word association, rhyme, and word fragment norms. Available at: http://w3.usf.edu/FreeAssociation/ Accessed February 2, 2009.

Nielsen, J. (1997). Be succinct! (writing for the Web). *Nielsen's Alertbox for March 15, 1997*. Available at: http://www.useit.com/alertbox/9703b.html Accessed October 24, 2009.

Nielsen, J. (2006). F-shaped pattern for reading web content. *Nielsen's Alertbox for April 17, 2006*. Available at: http://www.useit.com/alertbox/reading_pattern.html Accessed July 24, 2006.

Pan, B., Hembrooke, H. A., Gay, G. K., Granka, L. A., Feusner, M. K., & Newman, J. K. (2004). The determinants of web page viewing behavior: an eye-tracking study. In A. T. Duchowski & R. Vertegaal (Eds.), *Etra 2004: Proceedings of the eye tracking research & application symposium* (pp. 147–154). New York, NY, USA: ACM.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology*, *14* (3), 534–552.

Payne, J. W., Braunstein, M. L., & Carroll, J. S. (1978). Exploring pre-decisional behavior: Alternative approach to decision research. *Organizational Behavior and Human Decision Processes*, *22* (1), 17–44.

Pearson, R., & Van Schaik, P. (2003). The effect of spatial layout of and link colour in web pages on performance in a visual search task and an interactive search task. *International Journal of Human-Computer Studies*, *59* (3), 327–353.

Peavler, P. W. (1974). Pupil size, information overload, and performance differences. *Psychophysiology*, *11* (5), 559–566.

Pincombe, B. M. (2004). *Comparison of human and LSA judgements of pairwise document similarities for a news corpus* (Tech. Rep. No. DSTO-RR-0278). Adelaide, Australia: Australian Defense Science and Techology Organisation (DSTO), Intelligence, Surveillance and Reconnaissance Division. Available at: http://hdl.handle.net/1947/3334 Accessed April 15, 2008.

Pirolli, P. (1997). Computational models of information scent-following in a very large browsable text collection. In S. Pemberton (Ed.), *CHI '97: Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 3–10). New York, NY: ACM.

Pirolli, P. (2005). Rational analyses of information foraging on the web. *Cognitive Science*, *29* (3), 343–373.

Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review*, *106* (4), 643–675.

Pirolli, P., & Fu, W.-T. (2003). SNIF-ACT: A model of information foraging on the World Wide Web. In P. Brusilovsky, A. Corbett, & F. de Rosis (Eds.), *User modeling 2003: 9th international conference on user modeling* (pp. 45–54). Berlin, Germany: Springer-Verlag.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124* (3), 372–422.

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*, *11* , 95–113.

Richer, F., Silverman, C., & Beatty, J. (1983). Response selection and initiation in speeded reactions: a pupillometric analysis. *Journl of Experimental Psychology: Human Perception and Performance*, *9* (3), 360–370.

Rigutti, S., & Gerbino, W. (2004). Navigating within a web site: the WebStep Model. In M. C. Lovett, C. D. Schunn, C. Lebiere, & P. Munro (Eds.), *Proceedings of the sixth international conference on cognitive modeling* (pp. 378–379). Mahwah, NJ: Lawrence Erlbaum Associates.

Russo, J. E., & Dosher, B. A. (1983). Strategies for multiattribute binary choice. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *9* (4), 676–696.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18* (11), 613–620.

Schaeffer, T., Ferguson, J. B., Klein, J. A., & Rawson, E. B. (1968). Pupillary responses during mental activities. *Psychonomic Science*, *12* (4), 137–138.

Schluroff, M. (1982). Pupil responses to grammatical complexity of sentences. *Science*, *17* (1), 133–145.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6* (2), 461–464.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34* (1), 1–47.

Shashua, A., & Hazan, T. (2005). Non-negative tensor factorization with applications to statistics and computer vision. In L. De Raedt & S. Wrobel (Eds.), *Proceedings of the 22nd international conference on machine learning* (pp. 792–799). New York, NY: ACM Press.

Sokolov, Y. N. (1963). *Perception and the conditioned reflex*. Oxford, UK: Pergamon Press.

Spinks, J. A., & Siddle, D. (1972). The functional significance of the orienting response. In D. Siddle (Ed.), *Orienting and habituation: Perspectives in human research* (pp. 237–314). Chichester, UK: Wiley.

Spool, J. M., Schroeder, W., Scanlon, T., & Snyder, C. (1998). Web sites that work: designing with your eyes open. In C. M. Karat, A. Lund, J. Coutaz, & J. Karat (Eds.), *CHI '98: CHI 98 conference summary on Human Factors in computing systems* (pp. 147–148). New York, NY, USA: ACM.

Stanners, R. F., Coulter, I. M., Sweet, A. W., & Murphy, P. (1979). The pupillary response as an indicator of arousal and cognition. *Motivation and Emotion*, *3* (4), 319–340.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87* (2), 245–251.

Steinhauer, S. R., Condray, R., & Kasparek, A. (2000). Cognitive modulation of midbrain function: task-induced reduction of the pupillary light reflex. *International Journal of Psychophysiology*, *39* (1), 21–30.

Stone, B., & Dennis, S. (2007). Using LSA Semantic Fields to predict eye movement on web pages. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th annual conference of the Cognitive Society* (pp. 665–670). Mahwah, NJ: Lawrence Erlbaum Associates.

Stone, B., Dennis, S., & Kwantes, P. J. (in press). Comparing methods for paragraph similarity analysis. *Topics in Cognitive Science*.

Stone, B., Lee, M., Dennis, S., & Nettelbeck, T. (2004). Pupil size and mental load. *1st Adelaide Mental Life Conference. Available at: http://www.psychology.adelaide.edu.au/cognition/aml/ Accessed April 2, 2009*.

Toffler, A. (1973). *Future shock*. London, UK: Pan.

Turney, P. (2001). Mining the web for synonyms: PMI versus LSA on TOEFL. In L. De Raedt & P. A. Flach (Eds.), *ECML 2001: 12th European conference on machine learning* (pp. 491–502). Berlin, Germany: Springer.

Veitch, J. A., & McColl, S. L. (1995). Modulation of fluorescent light: Flicker rate and light source effects on visual performance and visual comfort. *Lighting Research and Technology*, *27* (4), 243–256.

Ward, R. D., & Marsden, P. H. (2003). Physiological responses to different WEB page designs. *International Journal of Human-Computer Studies*, *59* (1–2), 199–212.

Wu, S.-C., & Miller, C. S. (2007). Preliminary evidence for top-down and bottom-up processes in web

search navigation. In M. B. Rosson & D. J. Gilmore (Eds.), *CHI '07: CHI '07 extended abstracts on human factors in computing systems* (pp. 2765–2770).

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In J. Callan, D. Hawking, A. Smeaton, & C. Clarke (Eds.), *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '03)* (pp. 267–273). New York, NY: ACM Press.

Zelikovitz, S., & Kogan, M. (2006). Using web searches on important words to create background sets for LSI classification. In G. Sutcliffe & R. Goebel (Eds.), *Proceedings of the 19th international FLAIRS conference* (pp. 598–603). Menlo Park, CA: AAAI Press.

# Appendix A

## A. Paper 1: Statement of contributions

# Chapter 3: Pupil Size and Mental Load (2004)

Stone, B., Lee, M., Dennis, S., & Nettelbeck, T.

School of Psychology, University of Adelaide

1st Adelaide Mental Life Conference.

Statement of Contributions

Michael Lee (Co-author)

I was the supervisor (adviser) for the research program that lead to this publication. In terms of conceptualization of the program, there was extensive and ongoing collaboration between Mr. Stone and me in developing the direction of the research. The realization of the program, specifically, development of software, data collection and analyzes, were the work of Mr. Stone. I had an advisory role with respect to the direction and specifics of the data analyzes.

Mr. Stone was responsible for writing this paper; my role was to comment on drafts, make suggestions on the presentation of material in the paper, and to provide editorial input. I also provided advice on responding to comments by the article reviewers and editor.

I hereby give my permission for this paper to be incorporated in Mr. Stone's submission for the degree of Ph.D. in the University of Adelaide.

SIGNED: DATE: . . . 27 January 2010 . . . .

# Chapter 3: Pupil Size and Mental Load (2004)

Stone, B., Lee, M., Dennis, S., & Nettelbeck, T.

School of Psychology, University of Adelaide

1st Adelaide Mental Life Conference.

Statement of Contributions

Simon Dennis (Co-author)

I was a supervisor (adviser) for the research program that lead to this publication. The realization of the program, specifically, development of software, data collection and analyzes, were the work of Mr. Stone. The realization of the program, specifically, development of software, data collection and analyzes, were the work of Mr. Stone.

Mr. Stone was responsible for writing this paper; my role was to comment on drafts, make suggestions on the presentation of material in the paper, and to provide editorial input.

I hereby give my permission for this paper to be incorporated in Mr. Stone's submission for the degree of Ph.D. in the University of Adelaide.

SIGNED: DATE: 1/27/10

# Chapter 3: Pupil Size and Mental Load (2004)

Stone, B., Lee, M., Dennis, S., & Nettelbeck, T.

School of Psychology, University of Adelaide

1st Adelaide Mental Life Conference.

Statement of Contributions

Ted Nettelbeck (Co-author)

I was a supervisor (adviser) for the research program that lead to this publication. The realization of the program, specifically, development of software, data collection and analyzes, were the work of Mr. Stone. The realization of the program, specifically, development of software, data collection and analyzes, were the work of Mr. Stone.

Mr. Stone was responsible for writing this paper; my role was to comment on drafts, make suggestions on the presentation of material in the paper, and to provide editorial input.

I hereby give my permission for this paper to be incorporated in Mr. Stone's submission for the degree of Ph.D. in the University of Adelaide.

SIGNED:

# Appendix B

## B. Paper 2: Statement of contributions

# Chapter 4: Using LSA Semantic Fields to Predict Eye Movement on Web Pages (2007)

Benjamin Stone and Simon Dennis

School of Psychology, University of Adelaide

Proceedings of the 29th annual conference of the Cognitive Society (pp. 665-670)

Statement of Contributions

Simon Dennis (Co-author)

I was the supervisor (adviser) for the research program that lead to this publication. In terms of conceptualization of the program, there was extensive and ongoing collaboration between Mr. Stone and me in developing the direction of the research. The realization of the program, specifically, development of software, data collection, data modeling and analyzes, were the work of Mr. Stone. I had an advisory role with respect to selection of models used and on the direction and specifics of the data analyzes.

Mr. Stone was responsible for writing this paper; my role was to comment on drafts, make suggestions on the presentation of material in the paper, and to provide editorial input. I also provided advice on responding to comments by the article reviewers and editor.

I hereby give my permission for this paper to be incorporated in Mr. Stones submission for the degree of Ph.D. in the University of Adelaide.

SIGNED: DATE: 1/27/10

# Appendix C

## C. Paper 3: Statement of contributions

# Chapter 5: Comparing Methods for Single Paragraph Similarity Analysis (in press)

Benjamin Stone

School of Psychology, The University of Adelaide

Simon Dennis

Department of Psychology, Ohio State University

Peter J. Kwantes

Defence Research and Development Canada (Toronto)

in press, Topics in Cognitive Science.

Statement of Contributions

Simon Dennis (Co-author)

I was the supervisor (adviser) for the research program that lead to this publication. In terms of conceptualization of the program, there was extensive and ongoing collaboration between Mr. Stone and me in developing the direction of the research. The realization of the program, specifically, data modeling and and analyzes, were the work of Mr. Stone. I had an advisory role with respect to selection of models used and on the direction and specifics of the data analyzes.

Mr. Stone was responsible for writing this paper; my role was to comment on drafts, make suggestions on the presentation of material in the paper, and to provide editorial input. I also provided advice on responding to comments by the journal reviewers and editor.

I hereby give my permission for this paper to be incorporated in Mr. Stone's submission for the degree of Ph.D. in the University of Adelaide.

SIGNED: DATE: 1/27/10

# Chapter 5: Comparing Methods for Single Paragraph Similarity Analysis (in press)

Benjamin Stone

School of Psychology, The University of Adelaide

Simon Dennis

Department of Psychology, Ohio State University

Peter J. Kwantes

Defence Research and Development Canada (Toronto)

in press, Topics in Cognitive Science.

Statement of Contributions

Peter Kwantes (Co-author)

Empirical data referred to in the paper as the WENN dataset were collect by me and supplied to Mr Stone for this research. Also, the both the WENN and the Toronto Star corpora were collect by me and supplied to Mr Stone. The realization of the program, specifically, data modeling and analyzes, were the work of Mr. Stone.

Mr. Stone was responsible for writing this paper; my role was to comment on drafts, make suggestions on the presentation of material in the paper, and to provide editorial input.

I hereby give my permission for this paper to be incorporated in Mr. Stone's submission for the degree of Ph.D. in the University of Adelaide.

SIGNED: DATE: Jan 27/10

# Appendix D

## D. Paper 4: Statement of contributions

# Chapter 6: Semantic Models and Corpora Choice when using Semantic Fields to Predict Eye Movement on Web pages (submitted)

Benjamin Stone

School of Psychology, The University of Adelaide

Simon Dennis

Department of Psychology, Ohio State University

submitted, International Journal of Human-Computer Studies.

Statement of Contributions

Simon Dennis (Co-author)

I was the supervisor (adviser) for the research program that lead to this publication. In terms of conceptualization of the program, there was extensive and ongoing collaboration between Mr. Stone and me in developing the direction of the research. The realization of the program, specifically, development of software, data collection, data modeling and analyzes, were the work of Mr. Stone. I had an advisory role with respect to selection of models used and on the direction and specifics of the data analyzes.

Mr. Stone was responsible for writing this paper; my role was to comment on drafts, make suggestions on the presentation of material in the paper, and to provide editorial input. I will also provided advice on responding to comments by the journal reviewers and editor.

I hereby give my permission for this paper to be incorporated in Mr. Stone's submission for the degree of Ph.D. in the University of Adelaide.

1/27/10

# Appendix E

## E. Appendices from Paper 3 (Chapter 5)

*E.1. Examples of similar and dissimilar paragraphs as rated by humans for the WENN dataset*

*Similar paragraphs*

Paragraph 1: The woman accused of stalking Catherine Zeta-Jones will plead not guilty when she appears in court in two weeks time. Dawnette Knight is due to stand trial on November 10 on one charge of stalking and 24 charges of making criminal threats to the Chicago actress. Knight is currently being held in police custody, but her lawyers are appealing for bail, claiming she is only facing criminal proceedings because the case involves celebrities. Her lawyer Richard Herman says, "I'm afraid that we've gone much too far in something that's just greatly overblown. She's been in long enough. We all know that she's no threat to anyone." Welsh-born star Zeta-Jones, 35, who is married to actor Michael Douglas, said the "satanic threats" - which she received while she was filming Ocean's Twelve in Amsterdam last year - left her on the verge of a nervous breakdown.

Paragraph 2: The American woman accused of stalking Catherine Zeta-Jones has written a letter of apology to the star. Dawnette Knight - who is being held on $1 million bail - has penned a note to the Oscar-winning actress, and her father-in-law Kirk Douglas, apologizing for any "distress" she may have caused. Knight, was arrested at her Beverly Hills home on June 3 and charged with one count of stalking and 24 counts of criminal threats. She admits she is obsessed with Zeta-Jones' husband, Michael Douglas. Zeta-Jones' lawyer Richard Herman released the letter which reads: "I was a confused young woman infatuated with Michael Douglas and have not rational explanations for my actions." Her letter continues to explain she would "have never done anyone any harm and would never harm anyone". She finishes: "It would be a wonderful good deed if you would all forgive me so that I can go back to college to finish my studies in child psychology." Knight, 32, is being held in custody and

could receive a prison sentence of up to 19 years if convicted.

*Dissimilar paragraphs*

Paragraph 1: Movie superstar Tom Cruise has become the highest earning actor in Hollywood history after signing a deal that could earn him a staggering $360 million for his role in War Of The Worlds. Rather than agree a set fee for his part in the Steven Spielberg-directed epic, Cruise will earn 10 per cent of the film's box office takings plus a share of profits from DVDs, video games and toys. Experts predict the film - based on HG Wells' classic novel about a Martian attack - could make $1.8 billion at the cinema alone, of which Cruise's share would be an incredible $180 million. And, if he stars in the two planned sequels, Cruise's earnings will double at least. A Hollywood source says, "No expense will be spared. Spielberg wants to make it the film of the decade - the one that everyone talks about and rushes to see."

Paragraph 2: Superstar couple Victoria Beckham and David Beckham are desperate to add another child to their family in a bid to repair the damage done to their marriage by details of the soccer ace's alleged infidelity. The sexy pair recently canceled a planned promotional trip to America in favor of a two-week break in Morocco, and British newspaper The Sun reports they're using their intimate spell in exotic capital Marrakech to try for another baby. Friends of the couple claim they're ideally hoping to welcome a baby girl into the world to give sons Brooklyn, five, and 21-month-old Romeo a younger sister. The Beckhams' marriage became the subject of intense scrutiny earlier this year when David's ex-personal assistant Rebecca Loos revealed she'd enjoyed a steamy affair with the Real Madrid player. A source says, "They have talked about having more children and would be thrilled if they had a little girl. There's nothing either of them feel is more important than their kids - and David simply adores them. A new baby would be a great way for them to put their troubles behind them and start a new life together in Spain."

*E.2. Examples of similar and dissimilar paragraphs as rated by humans for the Lee dataset*

*Similar paragraphs*

Paragraph 1: Nigerian President Olusegun Obasanjo said he will weep if a single mother sentenced to death by stoning for having a child out of wedlock is killed, but added he has faith the court system will overturn her sentence. Obasanjo's comments late Saturday appeared to confirm he would not intervene directly in the case, despite an international outcry.

Paragraph 2: An Islamic high court in northern Nigeria rejected an appeal today by a single mother sentenced to be stoned to death for having sex out of wedlock. Clutching her baby daughter, Amina Lawal burst into tears as the judge delivered the ruling. Lawal, 30, was first sentenced in March after giving birth to a daughter more than nine months after divorcing.

*Dissimilar paragraphs*

Paragraph 1: Very few women have been appointed to head independent schools, thwarting efforts to show women as good leaders, according to the Victorian Independent Education Union. Although they make up two-thirds of teaching staff, women hold only one-third of principal positions, the union's general secretary, Tony Keenan, said. He believed some women were reluctant to become principals because of the long hours and the nature of the work. But in other cases they were shut out of the top position because of perceptions about their ability to lead and provide discipline.

Paragraph 2: Beijing has abruptly withdrawn a new car registration system after drivers demonstrated "an unhealthy fixation" with symbols of Western military and industrial strength - such as FBI and 007. Senior officials have been infuriated by a popular demonstration of interest in American institutions such as the FBI. Particularly galling was one man's choice of TMD, which stands for Theatre Missile Defence, a US-designed missile system that is regularly vilified by Chinese propaganda channels.

*E.3. Standard stop-list*

The stop-list used in this research prior to the removal of single numbers and letters.

a about above across after afterwards again against all almost alone along already also although always am among amongst amoungst amount an and another any anyhow anyone anything anyway anywhere are around as at back be became because become becomes becoming been before beforehand behind being below beside besides between beyond bill both bottom but by call can cannot cant co computer con could couldnt cry de describe detail do done down due during each eg eight either eleven else elsewhere empty enough etc even ever every everyone everything everywhere except few fifteen fify fill find fire first five for former formerly forty found four from front full further get give go had has hasnt have he hence her here hereafter hereby herein hereupon hers herself him himself his how however hundred i ie if in inc indeed interest into is it its itself keep last latter latterly least less ltd made many may me meanwhile might mill mine more moreover most mostly move much must my myself name namely neither never nevertheless next nine no nobody none noone nor not nothing now nowhere of off often on once one only onto or other others otherwise our ours ourselves out over own part per perhaps please put rather re same see seem seemed seeming seems serious several she should show side since sincere six sixty so some somehow someone something sometime sometimes somewhere still such system take ten than that the their them themselves then thence there thereafter thereby therefore therein thereupon these they thick thin third this those though three through throughout thru thus to together too top toward towards twelve twenty two un under until up upon us very via was we well were what whatever when whence whenever where whereafter whereas whereby wherein whereupon wherever whether which while whither who whoever whole whom whose why will with within without would yet you your yours yourself yourselves

*E.4. Corpora Parameters*

Parameters for each corpus used in this research are contain in Table E.4.1.

Table E.4.1: Corpus parameters for the Toronto Star corpus, WENN corpus, and sub-corpora drawn from Wikipedia (1000 and 10000 documents) for both WENN and Lee datasets.

| | Docs (D) | Words (W) | W/D | Unique W (UW) | UW/D |
|---|---|---|---|---|---|
| WENN | 12,787 | 957,806 | 74.91 | 22,915 | 1.79 |
| WIKI-WENN | 1,000 | 2,420,436 | 2,420.44 | 48,508 | 48.51 |
| WIKI-WENN | 10,000 | 26,983,256 | 2,698.33 | 180,225 | 18.02 |
| Toronto Star | 55,021 | 14,070,668 | 255.73 | 96,975 | 1.76 |
| WIKI-Lee | 1,000 | 2,267,287 | 2,267.29 | 34,285 | 34.29 |
| WIKI-Lee | 10,000 | 18,135,603 | 1,813.56 | 164,271 | 16.42 |

*E.5. Study One results tables*

*t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, with the WENN (see Table E.5.1) and Lee (see Table E.5.2) datasets generated by human raters in Study One.

*E.6. Stop-list used by Pincombe 2004*

Below is the stop-list used by Pincombe (2004) and Lee, Pincombe, and Welsh (2005). Also note that these researchers only included alphabetical characters (Pincombe, 2004, p. 14), therefore excluding both numbers and single letters from their corpora.

a about all also although am an and another any anybody anyhow anyone anything anywhere are as at b be become been being but by c can cannot could d did do does doing done e each eg either else et etc every ex f for from g h had has have having he hence her hers herself high him himself his how however i ie if in inc indeed is it its j k l ltd m many may me might more mr mrs ms must my myself n no nor not o of oh or otherwise ought our ours ourselves p per put q r re s self selves shall she should sl so some somehow such sup t than

Table E.5.1: *t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, with the human ratings contained in the WENN dataset. None of the models' performance significantly improved when dimensionality was increased (alpha 0.05). Significant decreases in performance would be indicated by negative values equal to or greater than 1.96. So, in no case was increased dimensionality associated with significant decrements to model performance.

| Model | 50 | 100 |
|---|---|---|
| LSA-100 | -0.31 | |
| LSA-300 | -0.66 | -0.76 |
| Topics-100 | 1.13 | |
| Topics-300 | 1.17 | 0.48 |
| Topics-JS-100 | 1.05 | |
| Topics-JS-300 | 0.34 | -0.31 |
| SpNMF-100 | 0.39 | |
| SpNMF-300 | -0.9 | -1.34 |

Table E.5.2: *t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, with human ratings contained in the Lee dataset. Significant decreases in performance would be indicated by negative values equal to or greater than 1.96. So, in no case was increased dimensionality associated with significant decrements to model performance.

| Model | 50 | 100 |
|---|---|---|
| LSA-100 | 2.29 | |
| LSA-300 | 2.30 | 2.28 |
| Topics-100 | -1.52 | |
| Topics-300 | 2.83 | 4.03 |
| Topics-JS-100 | 3.17 | |
| Topics-JS-300 | 2.30 | 1.60 |
| SpNMF-100 | 3.31 | |
| SpNMF-300 | 6.16 | 1.90 |

that the their theirs them themselves then there therefore these they this those though thus to u us v very via viz vs w was we were what whatever when whence whenever where whereafter whereas whereby wherein whereupon wherever whether which whichever while whither who whoever whole whom whose why will with within without would x y yes you your yours

yourself yourselves z

*E.7. Study Two result tables*

Correlations (*r*) between similarity assessments of human raters and those made using LSA, Topic Model (Topics), Topic Model with Jensen-Shannon equation (Topics-JS), SpNMF at 50, 100, and 300 dimensions, and both the Vectorspace and CSM models (see Table E.7.1). *t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, on the WENN (see Table E.7.2) and Lee (see Table E.7.3) datasets used in Study Two.

Table E.7.1: Correlations (*r*) between similarity assessments of human raters and those made using LSA, Topic Model (Topics), Topic Model with Jensen-Shannon equation (Topics-JS), SpNMF at 50, 100, and 300 dimensions, and also the Overlap, Vectorspace and CSM models. The ALL columns display correlations based on corpora that contain both numbers and single letters (as used in Study One), conversely the NN-NSL columns are based on corpora with No Numbers and No single Letters (NN-NSL). Correlations exclude Same-Same document comparisons.

| | WENN | | LEE | |
|---|---|---|---|---|
| Model: | ALL | NN-NSL | ALL | NN-NSL |
| Overlap | 0.43 | 0.62 | 0.48 | 0.53 |
| LSA-50 | 0.21 | 0.38 | 0.04 | 0.10 |
| LSA-100 | 0.20 | 0.41 | 0.05 | 0.11 |
| LSA-300 | 0.19 | 0.48 | 0.06 | 0.12 |
| Topics-50 | 0.01 | 0.22 | 0.02 | 0.07 |
| Topics-100 | 0.06 | 0.22 | 0.01 | 0.07 |
| Topics-300 | 0.08 | 0.25 | 0.06 | 0.07 |
| Topics-JS-50 | 0.11 | 0.26 | 0.10 | 0.15 |
| Topics-JS-100 | 0.12 | 0.29 | 0.11 | 0.17 |
| Topics-JS-300 | 0.12 | 0.28 | 0.13 | 0.17 |
| SpNMF-50 | 0.08 | 0.35 | 0.09 | 0.15 |
| SpNMF-100 | 0.09 | 0.37 | 0.13 | 0.16 |
| SpNMF-300 | 0.07 | 0.43 | 0.14 | 0.17 |
| Vectorspace | 0.17 | 0.41 | 0.10 | 0.20 |
| CSM | 0.26 | 0.16 | 0.15 | 0.17 |

Table E.7.2: *t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, with human ratings contained in the WENN dataset used in Study Two. All corpora have had single letters and numbers removed. Significant decreases in performance would be indicated by negative values equal to or greater than 1.96. In no case was increased dimensionality associated with significant decrements to model performance.

| Model | 50 | 100 |
|---|---|---|
| LSA-100 | 2.18 | |
| LSA-300 | 3.63 | 3.69 |
| Topics-100 | 0.16 | |
| Topics-300 | 0.58 | 1.12 |
| Topics-JS-100 | 1.22 | |
| Topics-JS-300 | 0.63 | -0.38 |
| SpNMF-100 | 1.87 | |
| SpNMF-300 | 2.2 | 1.7 |

Table E.7.3: *t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, with human ratings contained in the Lee dataset in Study Two. All corpora have had single letters and numbers removed. Significant decreases in performance would be indicated by negative values equal to or greater than 1.96. In no case was increased dimensionality associated with significant decrements to model performance.

| Model | 50 | 100 |
|---|---|---|
| LSA-100 | 3.16 | |
| LSA-300 | 2.98 | 2.53 |
| Topics-100 | 1.96 | |
| Topics-300 | 0.15 | -1.21 |
| Topics-JS-100 | 2.43 | |
| Topics-JS-300 | 1.33 | 0.16 |
| SpNMF-100 | 0.91 | |
| SpNMF-300 | 1.28 | 1.09 |

### *E.8. IMDB-based Lucene query for Wikipedia*

"naomi watts" OR "elevator" OR "fear" OR "liz* taylor" OR "eliz* taylor" OR "battling" OR "congestive" OR "heart" OR "failure" OR "whitney houston" OR "unhurt"

OR “crash” OR “zeta jones” OR “stalker” OR “apologizes” OR “trial” OR “courtney love” OR “drug” OR “case” OR “gary busey” OR “jailed” OR “late” OR “jessica simpson” OR “nick lachey” OR “split” OR “george clooney” OR “contemplates” OR “legal” OR “action” OR “kirsten dunst” OR “broke” OR “jake gyllenhaal” OR “heart” OR “pals” OR “jennifer lopez” OR “mother” OR “pines” OR “affleck” OR “lindsay lohan” OR “health” OR “battle” OR “michael jackson” OR “lawyers” OR “accuser” OR “undergo” OR “katherine zeta jones” OR “stalker” OR “plead” OR “guilty” OR “richard gere” OR “horseriding” OR “accident” OR “britney spears” OR “attacks” OR “paparazzi” OR “mary kate olsen” OR “comeback” OR “vmas” OR “tom cruise” OR “penelope cruz” OR “reunite” OR “dinner” OR “macaulay culkin” OR “arrested” OR “dog” OR “trip” OR “victoria beckham” OR “david beckham” OR “plan” OR “baby” OR “number” OR “demi moore” OR “pregnant” OR “posh” OR “victoria beckham” OR “flees” OR “france” OR “intruder” OR “incident” OR “ben affleck” OR “jennifer garner” OR “linked” OR “cameron diaz” OR “justin timberlake” OR “fight” OR “paparazzi” OR “tom cruise” OR “million” OR “war worlds” OR “newlyweds” OR “jessica simpson” OR “nick lachey” OR “look”

### *E.9. Lee-based Lucene query for Wikipedia*

(“australia*” and “democrats senator*” and “leader”) or (“amp” or “stock market”) or ((“mugabe” or “zimbabwe”) and “famine”) or (“alqaida” or “puk”) or ((“washington” or “us” or “usa” or “u s a” or “united states of america”) and (“georgian sovereignty” or “tbilisi”)) or ((“gay” or “homosexual” or “homo sexual”) and “discriminat*”) or (“saudi” and (“osama bin laden” or “bin laden” or “alqaida”)) or (“saddam hussein” or “abu nida”) ((“hunan” or “china”) and “flood”) or ((“warplanes” or “bombed”) and (“basra” or “iraq”)) or (“iraq” and “russia” and (“economic” or “cooperation”)) or (“saddam hussein” and (“weapons of mass destruction” or “wmd”)) or (“investigate” and (“taskforce” or “corruption”)) or ((“andrew bartlett” or “aden ridgeway” or “natasha stott despoja”) and “democrats”) or (“glass ceiling”

or “woman s rights” or “equal opportunity”) or (“war” and “iraq” and (“bush” or “us” or “usa” or “united states of america” or “u s a”)) or ((“beijing” or “chinese”) and “government”) or ((“africa*” or “malawi” or “mozambique” or “zambia” or “angola” or “swaziland” or “lesotho” or “zimbabwe”) and (“starv*” or “faminine”)) or ((“malawi” or “africa*”) and (“hiv” or “aids”)) or ((“un” or “united nation*” or “u n”) and “environment”) or ((“fatah revolutionary council” or “frc”) and “terror*”) or (“work” and “dole”) or (“anthrax” and “biowarfare”) or ((“china” or “chinese”) and “missile”) or (“death” and “stoning”) or (“death” and “stoned”) or (“warheads”) or ((“fbi” or “federal bureau investigation”) and “terrorism”) or (“tamal tiger*”) or (“crim*” and “voyeur*”) or (“crim*” and “video*”) or ((“australia*” or “tampa”) and (“refugee” or “asylum”)) or (“australia” and “democrat*”) or (“whale” and “rescue”) or ((“prince william” or “price harry” or “princess di*”) and “ken wharfe”) or ((“osama bin laden” or “bin laden”) and (“jihad” or “holy war”)) or (“johannesburg earth summit”) or (“mugabe” and “zimbabwe”) or ((“men s rights” or “mens rights”) and “movement”) or (“global warming” or (“environment*” and “degradation”)) or (“bird*” and “tag*” and “research*”) or ((“un” or “u n” or “united nations”) and “sustainable growth”) or ((“russia*” or “ussr” or “u s s r”) and “chinese worker*”) or (“australia*” and “tampa”) or (“batasuna”) or ((“river” or “water”) and “flood*”) or ((“europe*” and “palestin*”) and (“isreal*” or “jew*”)) or (“job” and (“cuts” or “retrench*”)) or ((“asylum” or “refugee”) and “australia*”)

### *E.10. Study Three result tables*

Human to model correlations when estimating similarity on the Wenn (see Table E.10.1) and Lee (see Table E.10.2) datasets, complex models using Wikipedia 1000 & 10000 document and domain-chosen corpora (without numbers or single letters – NN-NSL). Also, examples of the dimensions created by SpNMF on the 10000 document Wikipedia corpus generated for the Lee dataset (see Table E.10.3).

Table E.10.1: Human to model correlations when estimating paragraph similarity on the WENN dataset, complex models using Wiki(pedia) 1000 & Wiki 10000 document corpora and the WENN Corpus (NN-NSL). Correlations exclude Same-Same paragraph comparisons.

| Model | Corpus | *r* | Lower CI (95%) | Upper CI (95%) |
|---|---|---|---|---|
| Word Overlap | N/A | 0.62 | 0.53 | 0.69 |
| LSA | Wiki 1000 | 0.34 | 0.23 | 0.45 |
| LSA | Wiki 10000 | 0.26 | 0.14 | 0.37 |
| LSA | WENN | 0.48 | 0.38 | 0.57 |
| SpNMF | Wiki 1000 | 0.36 | 0.25 | 0.47 |
| SpNMF | Wiki 10000 | 0.39 | 0.28 | 0.49 |
| SpNMF | WENN | 0.43 | 0.33 | 0.53 |
| Topics | Wiki 1000 | 0.14 | 0.01 | 0.26 |
| Topics | Wiki 10000 | 0.24 | 0.13 | 0.36 |
| Topics | WENN | 0.25 | 0.13 | 0.36 |
| Topics-JS | Wiki 1000 | 0.27 | 0.15 | 0.38 |
| Topics-JS | Wiki 10000 | 0.25 | 0.13 | 0.36 |
| Topics-JS | WENN | 0.28 | 0.17 | 0.39 |
| Vectorspace | Wiki 1000 | 0.34 | 0.23 | 0.44 |
| Vectorspace | Wiki 10000 | 0.35 | 0.24 | 0.45 |
| Vectorspace | WENN | 0.41 | 0.30 | 0.51 |
| CSM | Wiki 1000 | 0.11 | -0.01 | 0.23 |
| CSM | Wiki 10000 | 0.12 | 0.00 | 0.24 |
| CSM | WENN | 0.16 | 0.04 | 0.28 |

Table E.10.2: Human to model correlations when estimating paragraph similarity on the Lee dataset, complex models using Wiki(pedia) 1000 & 10000 document corpora and the Toronto Star (NN-NSL) corpus. Correlations exclude Same-Same paragraph comparisons.

| Model | Corpus | *r* | Lower CI (95%) | Upper CI (95%) |
|---|---|---|---|---|
| Word Overlap | N/A | 0.53 | 0.49 | 0.57 |
| LSA | Wiki 1000 | 0.51 | 0.46 | 0.55 |
| LSA | Wiki 10000 | 0.39 | 0.34 | 0.43 |
| LSA | Toronto Star | 0.12 | 0.06 | 0.17 |
| SpNMF | Wiki 1000 | 0.53 | 0.49 | 0.57 |
| SpNMF | Wiki 10000 | 0.46 | 0.41 | 0.50 |
| SpNMF | Toronto Star | 0.17 | 0.11 | 0.22 |
| Topics | Wiki 1000 | 0.48 | 0.43 | 0.52 |
| Topics | Wiki 10000 | 0.43 | 0.38 | 0.47 |
| Topics | Toronto Star | 0.07 | 0.01 | 0.12 |
| Topics-JS | Wiki 1000 | 0.36 | 0.31 | 0.41 |
| Topics-JS | Wiki 10000 | 0.42 | 0.37 | 0.47 |
| Topics-JS | Toronto Star | 0.17 | 0.11 | 0.22 |
| Vectorspace | Wiki 1000 | 0.55 | 0.51 | 0.59 |
| Vectorspace | Wiki 10000 | 0.56 | 0.52 | 0.59 |
| Vectorspace | Toronto Star | 0.20 | 0.14 | 0.25 |
| CSM | Wiki 1000 | 0.08 | 0.02 | 0.13 |
| CSM | Wiki 10000 | 0.10 | 0.04 | 0.15 |
| CSM | Toronto Star | 0.17 | 0.12 | 0.22 |

Table E.10.3: Examples of dimensions created by SpNMF on the 10000 document Wikipedia corpus generated for the Lee dataset where document length has been truncated at 100 words.

| Dimension 1 | | Dimension 2 | | Dimension 3 | |
|---|---|---|---|---|---|
| pollution | 0.78 | weapons | 0.42 | al | 0.81 |
| climate | 0.21 | biological | 0.34 | qaeda | 0.21 |
| change | 0.14 | bwc | 0.25 | bin | 0.17 |
| global | 0.14 | warfare | 0.21 | laden | 0.14 |
| environmental | 0.13 | germ | 0.15 | itihaad | 0.10 |
| warming | 0.13 | toxin | 0.13 | osama | 0.10 |
| overuse | 0.12 | pathogen | 0.13 | qaida | 0.09 |
| waste | 0.12 | stockpiling | 0.13 | group | 0.09 |
| greenhouse | 0.11 | incapacitate | 0.13 | abd | 0.09 |
| ipcc | 0.10 | organism | 0.13 | fadl | 0.08 |
| contamination | 0.09 | adversary | 0.13 | islamic | 0.08 |
| fossil | 0.09 | agreement | 0.13 | islam | 0.08 |
| resources | 0.08 | virus | 0.12 | militant | 0.08 |
| overpopulation | 0.08 | employment | 0.12 | terrorist | 0.07 |
| fuels | 0.08 | disease | 0.12 | abu | 0.07 |
| water | 0.08 | outlawed | 0.11 | islamist | 0.07 |
| wmo | 0.08 | causing | 0.11 | sunni | 0.06 |
| conservation | 0.08 | devastating | 0.11 | nashiri | 0.06 |
| deforestation | 0.07 | impact | 0.11 | ahmed | 0.05 |
| issues | 0.07 | kill | 0.11 | ali | 0.05 |

### *E.11. Study Four results tables*

Comparisons of model performance when the 50 Lee dataset paragraphs are added to the Wikipedia 1000 (see Table E.11.1) and 10000 (see Table E.11.2) Wikipedia sub-corpora.

Table E.11.1: Comparison of models performance with standard Wikipedia 1000 corpora (Wiki 1000) and Wikipedia 1000 corpora including the 50 Lee paragraphs (Wiki 1050), using correlations between human and model estimates of paragraph similarity on the Lee dataset. Correlations exclude Same-Same paragraph comparisons. Significance tests were performed using Williams' T2 formula.

| Model | Wiki 1000 | Wiki 1050 | diff | t | p |
|---|---|---|---|---|---|
| LSA | 0.51 | 0.6 | 0.09 | 7.65 | $< 0.05$ |
| Vectorspace | 0.55 | 0.67 | 0.12 | 9.08 | $< 0.05$ |
| Topics | 0.48 | 0.49 | 0.01 | 3.02 | $< 0.05$ |
| Topics-JS | 0.36 | 0.36 | -0.01 | -1 | n.s. |
| SpNMF | 0.53 | 0.56 | 0.03 | 2.66 | $< 0.05$ |
| CSM | 0.08 | 0.08 | 0 | 6.51 | $< 0.05$ |

Table E.11.2: Comparison of models performance with standard Wikipedia 10000 corpora (Wiki 10000) and Wikipedia 10000 corpora including the 50 Lee paragraphs (Wiki 10050), using correlations between human and model estimates of paragraph similarity on the Lee dataset. Correlations exclude Same-Same paragraph comparisons. Significance tests were performed using Williams' T2 formula.

| Model | Wiki 10000 | Wiki 10050 | diff | t | p |
|---|---|---|---|---|---|
| LSA | 0.39 | 0.40 | 0.01 | 14.56 | $< 0.05$ |
| Vectorspace | 0.56 | 0.58 | 0.02 | 6.92 | $< 0.05$ |
| Topics | 0.43 | 0.49 | 0.06 | 7.18 | $< 0.05$ |
| Topics-JS | 0.42 | 0.40 | -0.02 | -2.41 | $< 0.05$ |
| SpNMF | 0.46 | 0.44 | -0.02 | -4.75 | $< 0.05$ |
| CSM | 0.10 | 0.10 | 0 | 6.57 | $< 0.05$ |

# Appendix F

## F. Appendices from Paper 4 (Chapter 6)

### *F.1. Goal pages with Semantic Field maps generated using Vectorspace and WIKI-WEB*

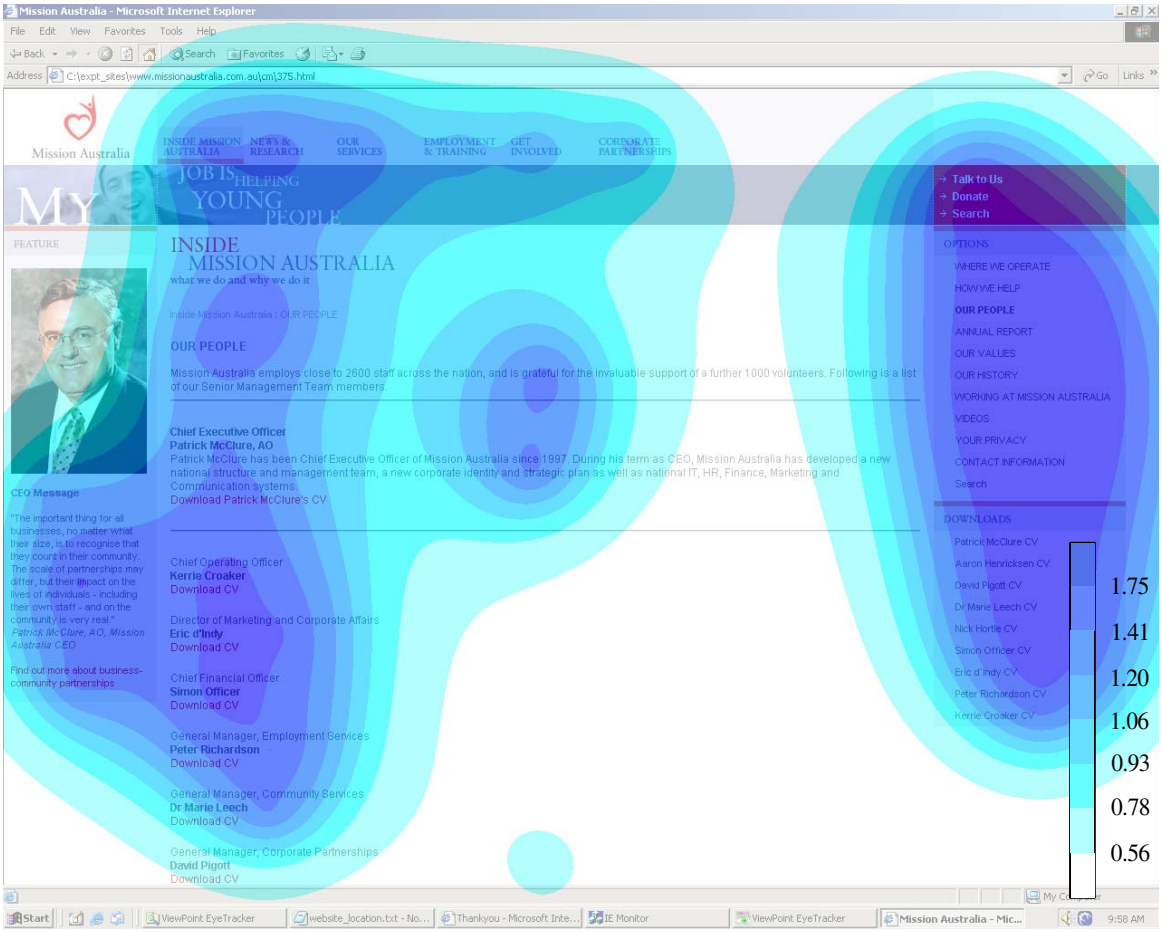



*Figure F.1.1.* Mission Australia - Task 1, "Who is currently the Chief Operating Officer of Mission Australia?" Areas of greater estimated goal-oriented information salience have darker colors in this heat map.

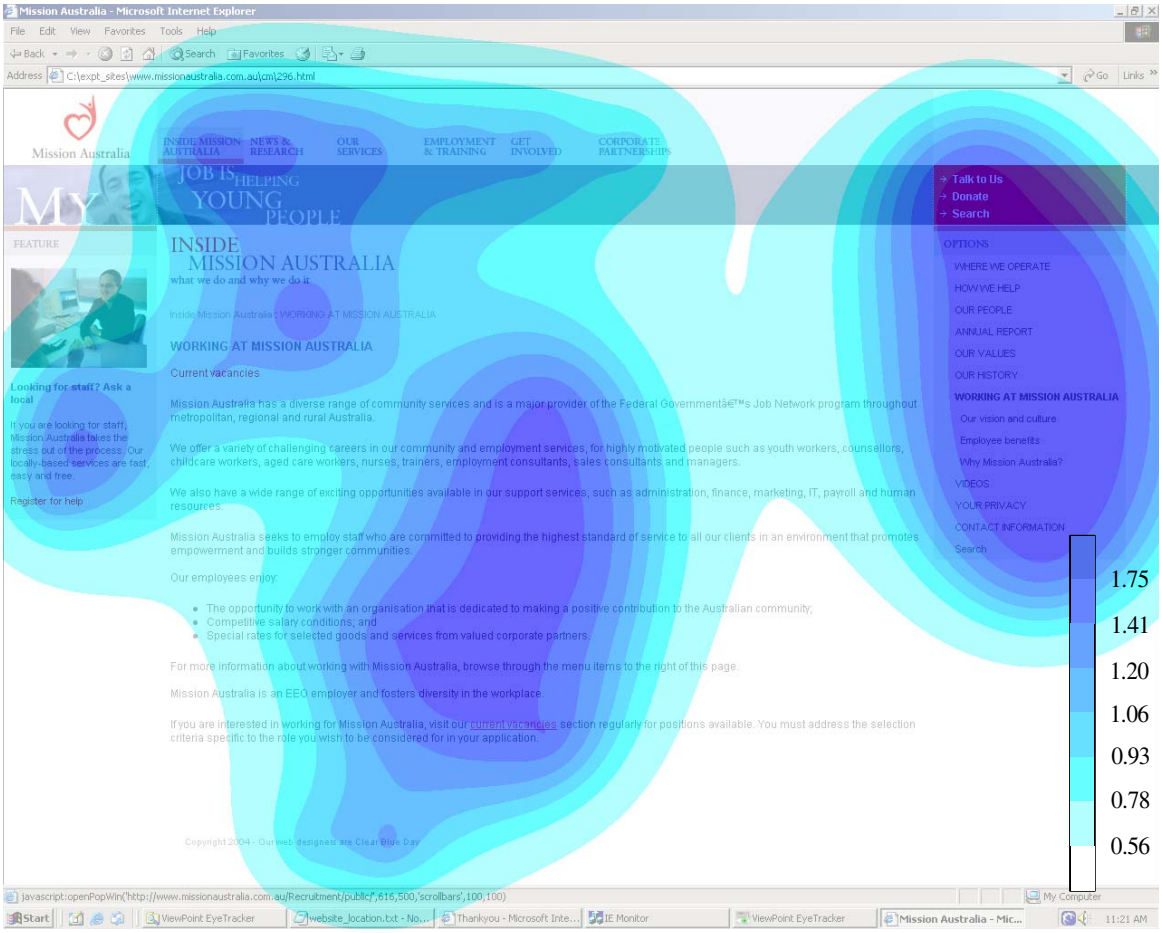



*Figure F.1.2.* Mission Australia - Task 2, "You are interested in working for Mission Australia. Search their Web site for the current job vacancies available at Mission Australia." Areas of greater estimated goal-oriented information salience have darker colors in this heat map.

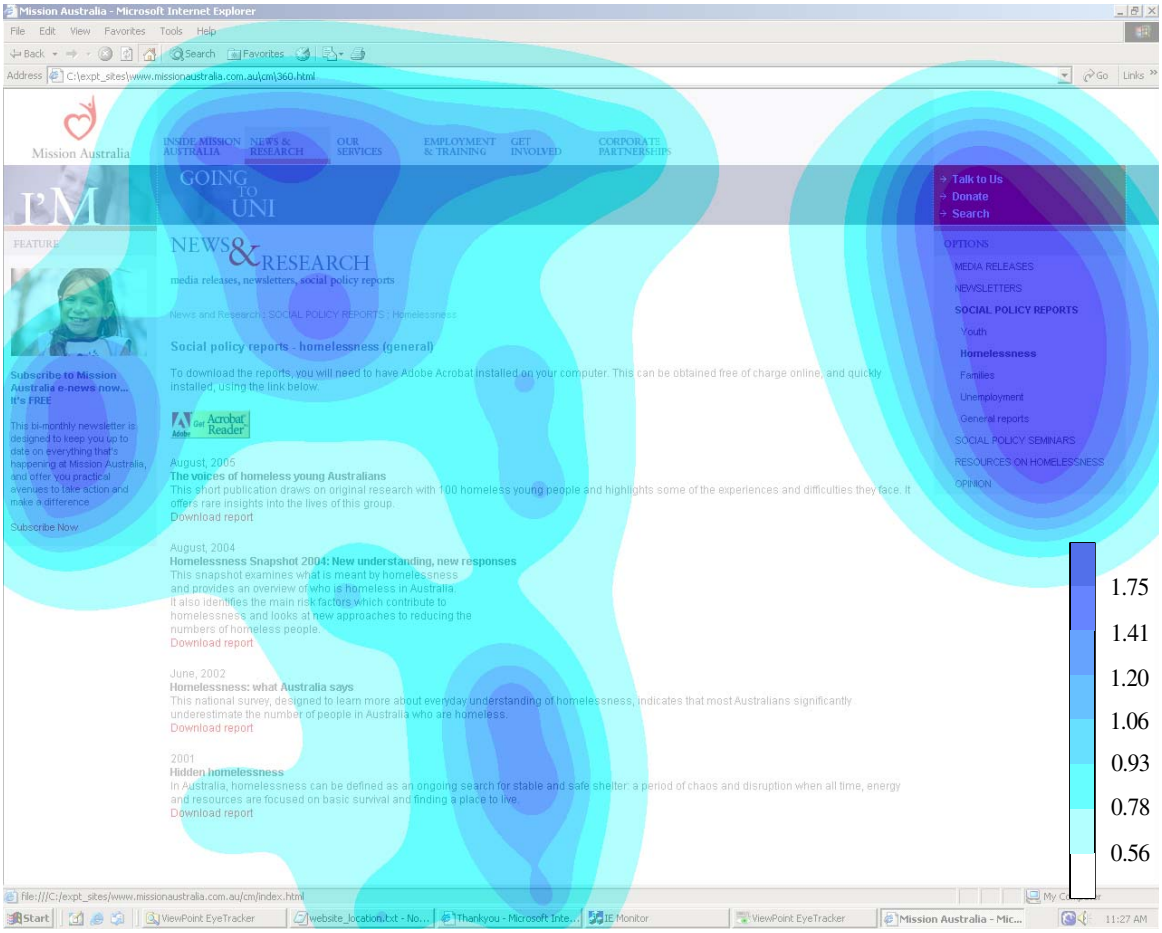



*Figure F.1.3.* Mission Australia - Task 3, "You are currently researching homelessness in young people and have heard that Mission Australia has recently published a report called 'The voices of homeless young Australians'. Search the Mission Australia Web site for this report into youth homelessness." Areas of greater estimated goal-oriented information salience have darker colors in this heat map.

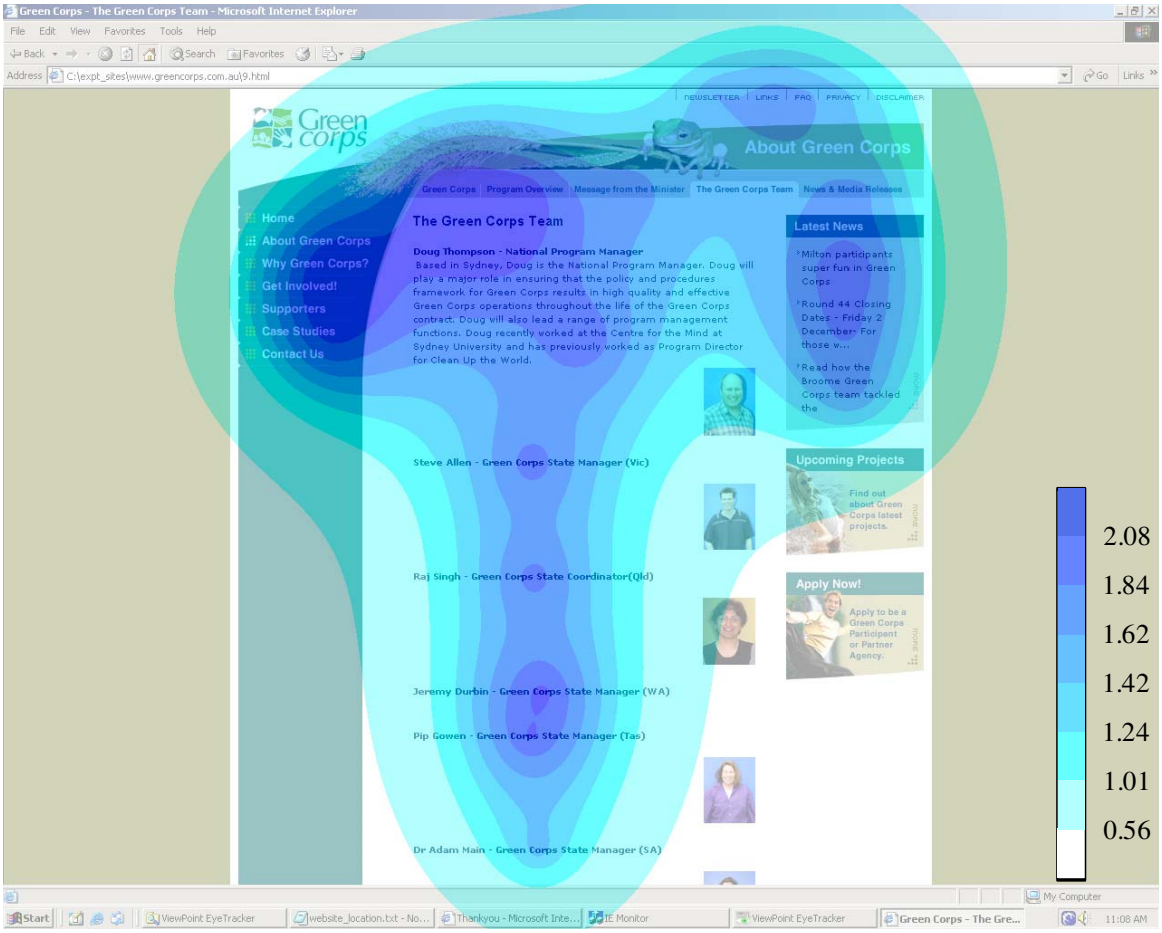



*Figure F.1.4.* Green Corps - Task 1, "You want to know more about Green Corps management. Find out who is the National Program Manager of Green Corps." Areas of greater estimated goal-oriented information salience have darker colors in this heat map.


*Figure F.1.5.* Green Corps - Task 2, "Find what environmental and heritage benefits are contributed by Green Corps." Areas of greater estimated goal-oriented information salience have darker colors in this heat map.

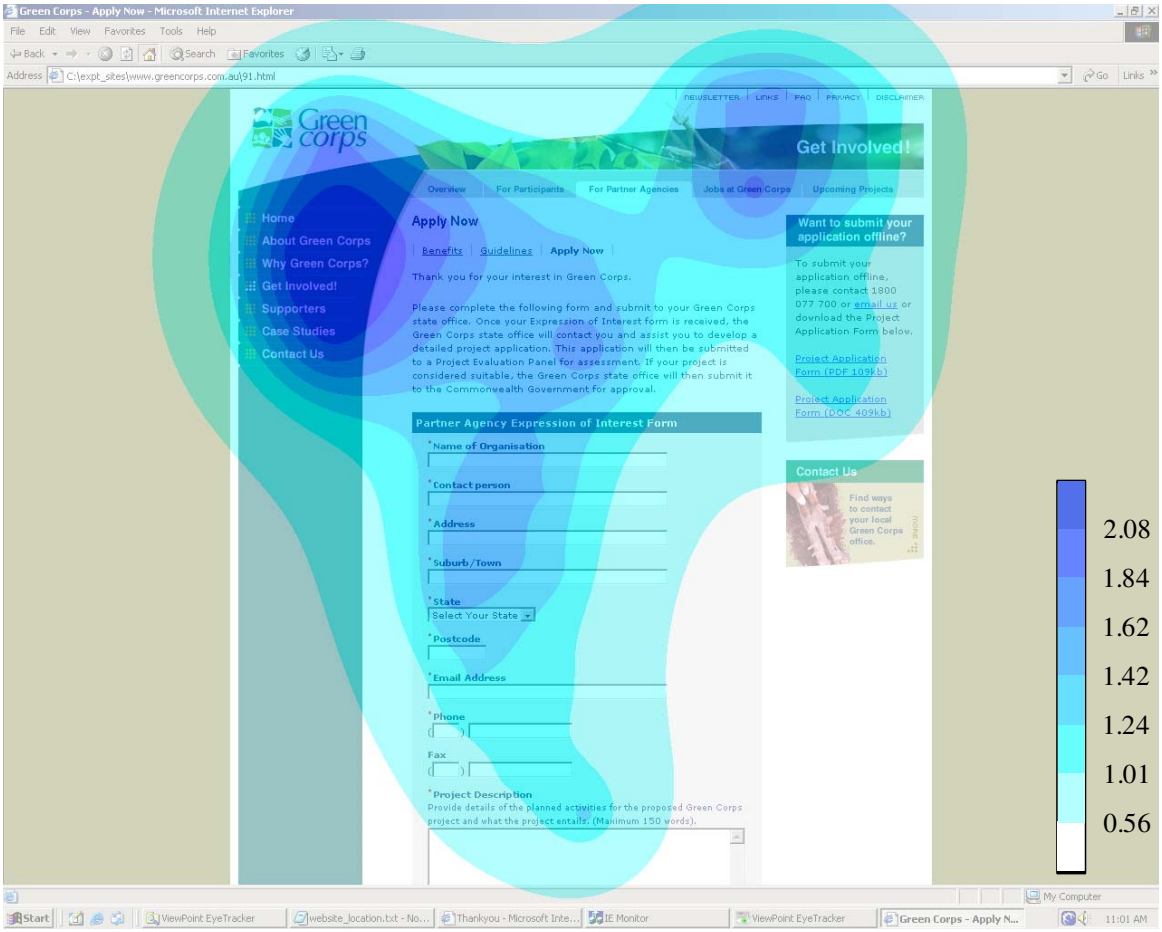



*Figure F.1.6.* Green Corps - Task 3, "Find the online Expression of Interest form to apply to become a Green Corps Partner Agency." Areas of greater estimated goal-oriented information salience have darker colors in this heat map.

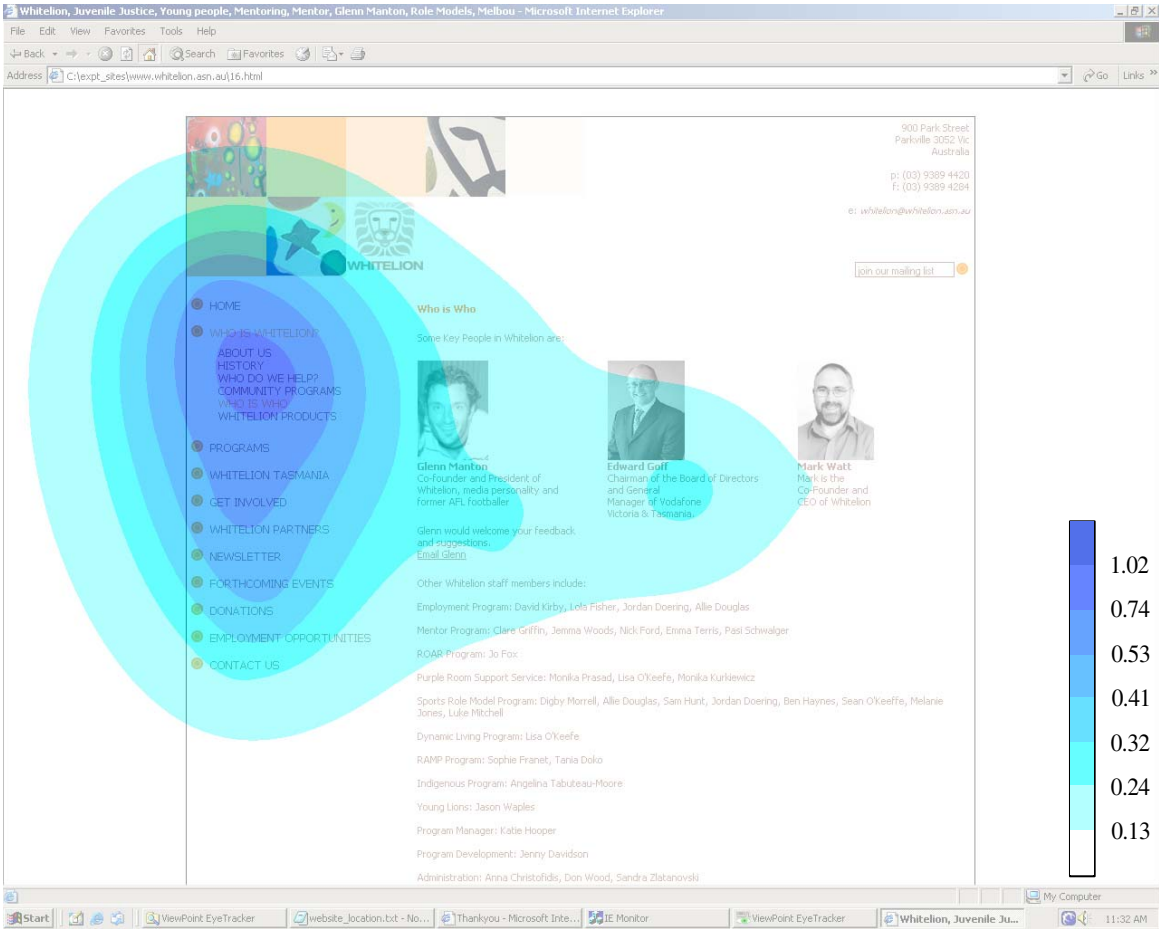



*Figure F.1.7.* White Lion - Task 1, "Find out who is the current President of White Lion." Areas of greater estimated goal-oriented information salience have darker colors in this heat map.

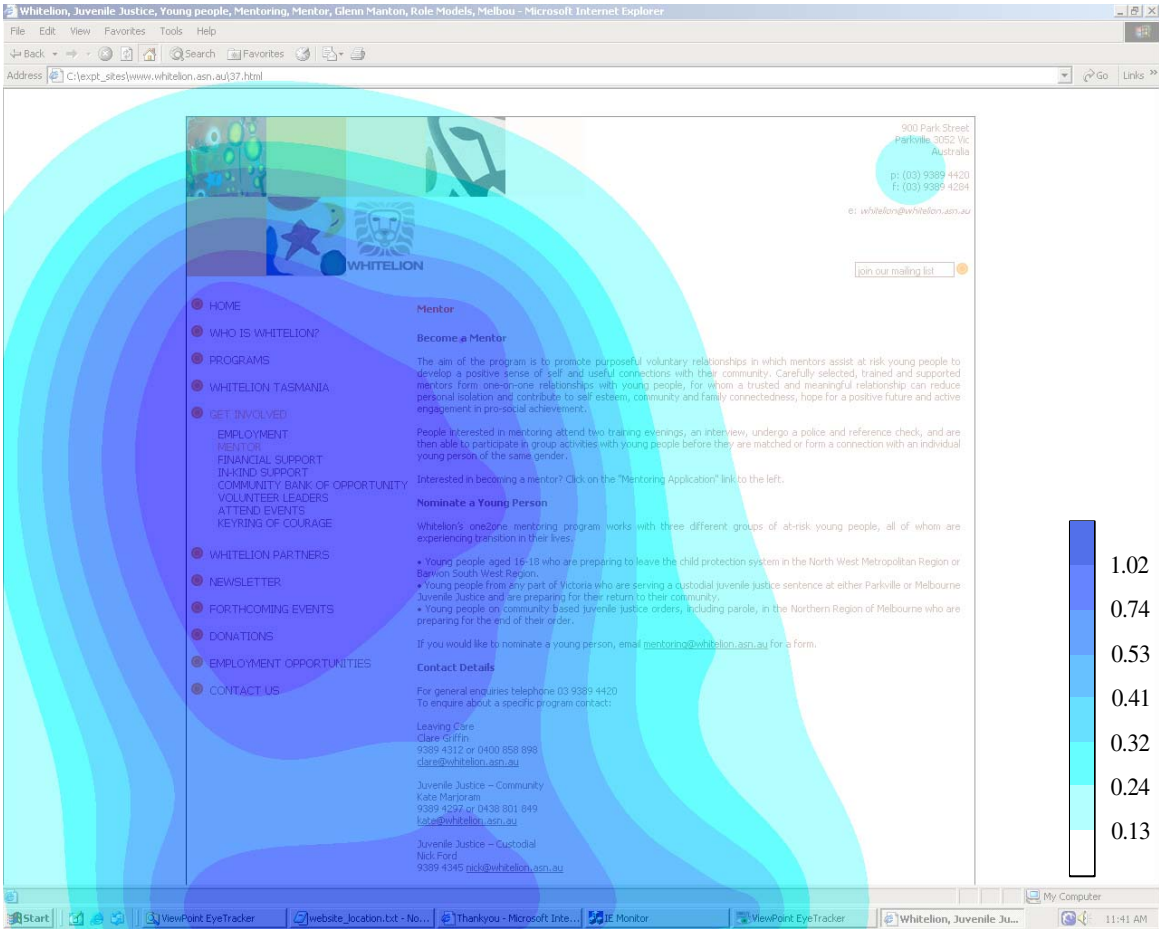



*Figure F.1.8.* White Lion - Task 2, "You are interested in becoming a mentor for young people. Find out how to become one of White Lions mentors." Areas of greater estimated goal-oriented information salience have darker colors in this heat map.

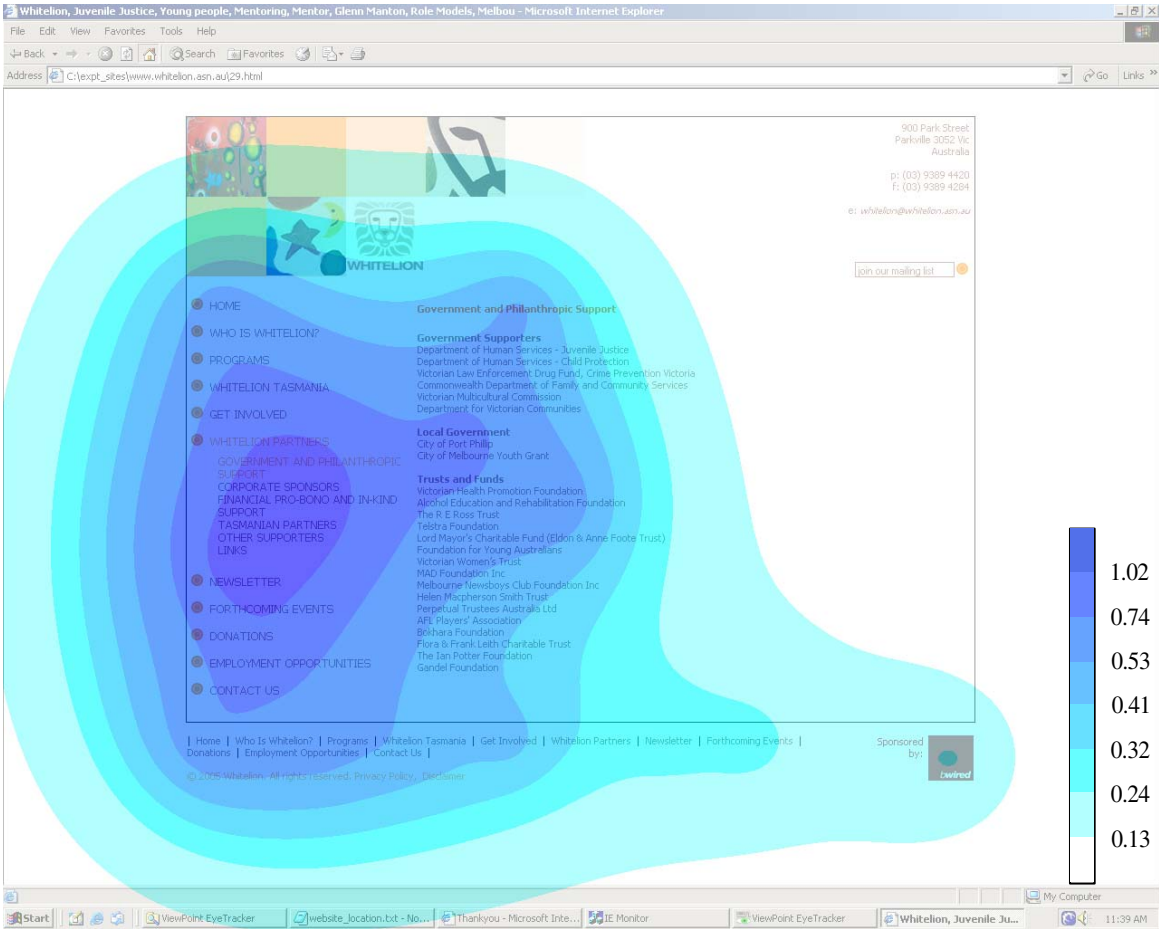



*Figure F.1.9.* White Lion - Task 3, "You are interested in financial viability of White Lion as a business. Find out which Government Departments are supporters of the White Lion organization." Areas of greater estimated goal-oriented information salience have darker colors in this heat map.

# Appendix G

## G. Paper 1 - Original Article

Stone, B., Lee, M., Dennis, S. & Nettelbeck, T. (2004). Pupil size and mental load. *1st Adelaide Mental Life Conference, Adelaide, S.A.*

# Appendix H

## H. Paper 2 - Original Article

Stone, B. & Dennis, S. (2007). Using LSA Semantic Fields to Predict Eye Movement on Web Pages.
In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 665-670). Austin, TX: Cognitive Science Society.

# Appendix I

## I. Paper 3 - Original Article

Running head: COMPARING METHODS FOR SINGLE PARAGRAPH SIMILARITY ANALYSIS

# Comparing Methods for Single Paragraph Similarity Analysis

Benjamin Stone

School of Psychology

The University of Adelaide

Simon Dennis

Department of Psychology

Ohio State University

Peter J. Kwantes

Defence Research and Development Canada (Toronto)

**Abstract**

The focus of this paper is two-fold. First, similarities generated from six semantic models were compared to human ratings of paragraph similarity on two datasets - 23 World Entertainment News Network paragraphs and 50 ABC newswire paragraphs. Contrary to findings on smaller textual units such as word associations (Grifths, Tenenbaum, & Steyvers, 2007), our results suggest that when single paragraphs are compared, simple non-reductive models (word overlap and vector space) can provide better similarity estimates than more complex models (LSA, Topic Model, SpNMF, and CSM). Second, various methods of corpus creation were explored to facilitate the semantic models' similarity estimates. Removing numeric and single characters, and also truncating document length improved performance. Automated construction of smaller Wikipedia-based corpora proved to be very effective even improving upon the performance of corpora that had been chosen for the domain. Model performance was further improved by augmenting corpora with dataset paragraphs.

## 1. Introduction

> The rate at which man [sic] has been storing up useful knowledge about himself and the universe has been spiralling upwards for 10,000 years.
>
> – (Toffler, 1973, p. 37)

Nearly four decades later, Toffler's remark is perhaps even more relevant in today's internet-driven world. 'Information overload' may be regarded as pervasive in many professions, and filtering strategies such as the summarization of text are commonplace. Government leaders and company executives make informed decisions based on briefs or short summaries of complex issues, provided by department managers who have in turn summarized longer reports written by their staff. In academia, the abstract is used to provide an overview of a paper's contents, so that time-pressed researchers can filter and absorb information related to their fields of study. In many areas it is important to be able to accurately judge the similarity between two or more paragraphs of information.

Sorting and extracting useful information from large collections of these types of summaries can prove both overwhelming and time consuming for humans. In an attempt to address this issue, semantic models have been successfully employed at these tasks. For example, Latent Semantic Analysis (LSA, Landauer, McNamara, Dennis, & Kintsch, 2007) has been used grade student essay scripts (Foltz, Laham, & Landauer, 1999). Similarly, the Topic Model has been used to extract scientific themes from abstracts contained in the Proceedings of the National Academy of Sciences (Griffiths & Steyvers, 2004). In a surveillance application, nonnegative matrix factorization has been applied to the large ENRON email dataset to extract topics or themes (Berry & Browne, 2005).

Other models such as the vector space model (henceforth called 'Vectorspace'), were originally designed to index (or order by relevance to a topic) large sets of documents (Salton, Wong, & Yang, 1975).

Semantic models have also been shown to reflect human knowledge in a variety of ways. LSA measures correlate highly with humans' scores on standard vocabulary and subject matter tests; mimic human word sorting and category judgments; simulate word-word and passage-word lexical priming data; and accurately estimate passage coherence (Landauer, McNamara, Dennis, & Kintsch, 2007). The Topic Model has proven adept at predicting human data on tasks including: free association, vocabulary tests, lexical decision, sentence reading and free recall (Griffiths, Tenenbaum, & Steyvers, 2007). Other models have been developed to reflect specific psychological processes. For example, the Constructed Semantics Model (CSM) was developed as a global-matching model of semantic memory derived from the MINERVA 2 architecture of episodic memory (Kwantes, 2005).

### *1.1. Different types of textual language unit*

When making similarity comparisons on textual stimuli with semantic models, several researchers have highlighted the need to delineate textual stimuli into different language units (Kireyev, 2008; Foltz, 2007; McNamara, Cai, & Louwerse, 2007; Landauer & Dumais, 1997). Past research has modeled human comparisons of similarity on four types of textual language units: words, sentences, single paragraphs and chapters or whole documents (Foltz, 2007).

*1.1.1. Word comparisons*. Griffiths et al. (2007) found that the Topic Model outperformed LSA on several tasks including word association and synonym

identification. Griffiths and colleagues compared performance by the Topic Model and LSA on a word association task using norms collected by Nelson, McEvoy, and Schreiber (1998). The study used 4471 of these words that were also found in an abridged[1] 37,651 document (26,243 word, 4,235,314 token) version of the Touchstone Applied Science Associates (TASA) corpus. Moreover, the TASA corpus was used as a knowledge base for both the Topic Model and LSA. Two measures were employed to assess the models' estimates of word association. The first measure assessed central tendency, focusing on the models' ability to rank word targets for each word cue. The other measure assessed the proficiency of each model's estimate of the most likely target response for each word cue. Griffiths and colleagues found that the Topic Model outperformed LSA on both of these performance measures. Furthermore, they reported that both models performed at levels better than chance and a simple word co-occurrence model.

In another study, Griffiths et al. (2007) compared the Topic model and LSA on a subset of the synonym section taken from the Test of English as a Foreign Language™(TOEFL®). The TOEFL was developed in 1963 by the National Council on the Testing of English as a Foreign Language, and is currently administered by the Educational Testing Service®[2]. The synonym portion of TOEFL offers four multiple choice options for each probe word, Griffiths and colleagues only included items in which all five words also appeared in the aforementioned abridged version of the TASA corpus. Similarity evaluations between the probes and possible synonyms, revealed that the Topics model (70.5%) answered more of the 44 questions correctly than LSA (63.6%). Furthermore, the Topic Model (0.46) predictions captured more of the variance found in the human responses than LSA (0.3).

The Topic model is a *generative model* that assesses the probability that words will

be assigned to a number of topics. One of the key benefits of this generative process is that it allows words to be assigned to more than one topic, thus accommodating the ambiguity associated with homographs (Griffiths et al., 2007). For example, using the Topic Model the word ‘mint’ may appear in a topic that contains the words ‘money’ and ‘coins’, and in another topic containing the words ‘herb’ and ‘plants’. Griffiths et al. (2007) argue that this attribute gives the Topic Model an advantage over models like LSA which represent meanings of words as individual points in undifferentiated Euclidean space (p. 219-220).

*1.1.2. Sentence comparisons*. McNamara et al. (2007) used several implementations of LSA to estimate the relatedness of sentences. The human judged similarity of these sentences decreased from paraphrases of target sentences, to sentences that were in the same passage as target sentences, to sentences that were selected from different passages to the target sentences. Likewise, comparing sentences using a standard implementation of LSA and the TASA corpus, these researchers found estimates of similarity were greatest for paraphrases, then same passage sentences, with different passage sentences judged least similar. When human estimates were correlated with the LSA estimates of sentences similarity, it was found that a version of LSA that emphasized frequent words in the LSA vectors best captured the human responses. Subsequently, using data collected in the McNamara et al. (2007) study, Kireyev (2008) found that LSA outperformed the Topic Model at this task.

*1.1.3. Single paragraph comparisons*. Lee, Pincombe, and Welsh (2005) examined similarity judgments made by Adelaide University students on 50 paragraphs that were collected from the Australian Broadcasting Corporations news mail service. These paragraphs ranged from 56 to 126 words in length, with a median length of 78.5 words. Lee and colleagues compared several models’ estimates of similarity to the

aforementioned human ratings. These models included word-based, n-gram and several LSA models. Using a knowledge base of 364 documents also drawn from the ABC news mail service, LSA under a global entropy function[3] was the best performing model, producing similarity ratings that correlated about 0.60 with human judgments in this study. LSA's result in this study was also consistent with the inter-rater correlation (approximately 0.605) calculated by these researchers.

More recently, Gabrilovich and Markovitch (2007) produced a substantially higher correlation with the human similarity judgments recorded for the Lee paragraphs (0.72) using the model they developed, Explicit Semantic Analysis (ESA). The ESA model uses Wikipedia as a knowledge base, treating Wikipedia documents as discrete human generated concepts that are ranked in relation to their similarity to a target text using a centroid-based classifier.

Kireyev (2008) used LSA and the Topic Model to estimate similarity of pairs of paragraphs taken from 3rd and 6th grade science textbooks. It was proposed that paragraphs that were adjacent, should be more similar than non-adjacent paragraphs. Difference scores were calculated between adjacent and non-adjacent paragraphs for both grade levels, with higher scores indicating better model performance. While it was not stated whether one model significantly outperformed the other at this task, on average LSA (0.75) scored higher on the 3rd Grade paragraphs than the Topic Model (0.49). However, there was little difference between the two models on the 6th Grade paragraphs (LSA 0.33, Topic Model 0.34).

*1.1.4. Chapters or whole document comparisons.* Martin and Foltz (2004) compared whole transcripts of team discourse to predict team performance during simulated reconnaissance missions. Sixty-seven mission transcripts were used to create

the researcher's corpus (UAV-Corpus). LSA was used to measure the similarity between transcripts of unknown missions to transcripts of missions where performance scores were known. To estimate the performance of a team based on their transcript using LSA, an average performance score was calculated from the 10 most similar transcripts found in the UAV-corpus. Performance scores estimated using LSA were found to correlate strongly (0.76) with the actual team performance scores.

Kireyev (2008) compared the similarity estimates of LSA and the Topic Model using 46 Wikipedia documents. These documents were drawn from six different categories: sports, animals, countries, sciences, religion, and disease. While both models correctly found more similarity between within-category documents than across-category documents, Kireyev (2008) concluded that LSA performed this task consistently better than the Topic Model.

### *1.2. The dual focus of this paper*

This paper describes the outcome of a systematic comparison of single paragraph similarities generated by six statistical semantic models to similarities generated by human participants. Paragraph complexity and length can vary widely. Therefore, for the purposes of this research, we define a paragraph as a self-contained section of 'news' media (such as a précis), presented in approximately 50 to 200 words.

There are two main themes that are explored in this paper. At one level it is an evaluation of the semantic models, in which their performance at estimating the similarity of single paragraph documents is compared against human judgments. As outlined above, past research has indicated that performance of some models is clearly better depending on which type of textual units were used as stimuli. For example, the Topic Model was shown to perform better than LSA in word association research, where the textual unit

was at the single word level. However, inherent difficulties such as homographs that affect models like LSA at the word unit level, may be less problematic for assessments made on larger textual units (sentences, paragraphs, and chapters or whole documents). These larger textual units contain concurrently presented words that may be less ambiguous and are thus able to compensate for a model's inability to accommodate homographic words (Landauer & Dumais, 1997; Choueka & Lusignan, 1985).

Research has indicated that LSA performs well at the paragraph level (Lee et al., 2005), but there are other models that may perform equally well if not better than LSA at this task. Therefore, in this research we compare six models' efficiency at the task of modeling human similarity judgments of single paragraph stimuli, the models examined were: word overlap, the Vectorspace model (Salton, Wong & Yang, 1975), Latent Semantic Analysis (LSA, Kintsch, McNamara, Dennis, & Landauer, 2007), the Topic Model (Griffiths & Steyvers, 2002, Blei, Ng, & Jordan, 2003), Sparse Nonnegative Matrix Factorization (SpNMF, Xu, Liu, & Gong, 2003), and the Constructed Semantics Model (CSM, Kwantes, 2005). Our evaluation of these models is tempered by factors such a model compilation speed, consistency of performance in relation to human judgments of document similarity, and intrinsic benefits such as producing interpretable dimensions.

At another level this paper explores the characteristics of the corpora or knowledge bases utilized by these models that may improve models' performance when approximating human similarity judgments. With the exception of the word overlap model, a good background knowledge base is essential to the models' performance. Past research has identified various aspects of corpus construction that affect the performance of the Pointwise Mutual Information (PMI) co-occurrence model on word-based tasks such as the TOEFL synonym test (Bullinaria & Levy, 2006). These factors included: the

size and shape of the context window, the number of vectors included in the word space, corpus size and corpus quality. To address this issue, we have evaluated aspects of corpus composition, preprocessing and document length in an attempt to produce suitable background corpora for the semantic models.

To this end, four studies are described in this paper that examine the semantic models' performance relative to human ratings of paragraph similarity. In the first study, semantic models use domain-chosen corpora to generate knowledge spaces on which they make evaluations of similarity for two datasets of paragraphs. Overall, the models performed poorly using these domain-chosen corpora when estimates were compared to those made by human assessors. In the second study, improvements in the models' performance were achieved by more thoroughly preprocessing the domain-chosen corpora to remove all instances of numeric and single alphabetical characters. In the third study, smaller targeted corpora (sub-corpora) constructed by querying a larger set of documents (Wikipedia[4]) were examined to assess whether they could produce sufficient performance to be generally useful (Zelikovitz & Kogan, 2006). In many applications the hand construction of corpora for a particular domain is not feasible, and so the ability to show a good match between human similarity evaluations and semantic models' evaluations of paragraph similarity using automated methods of corpus construction is a desirable outcome. Furthermore, document length of the essay-like Wikipedia articles was manipulated to produce better approximations of human judgment by the semantic models. Finally, in the fourth study, several of the models were found to produce better estimates of paragraph similarity when the dataset paragraphs were included in the models' backgrounding corpus.

## 2. Semantic models, human datasets and domain-chosen corpora

### *2.1. Semantic models*

The semantic models examined were word overlap, the Vectorspace model (Salton, Wong & Yang, 1975), Latent Semantic Analysis (LSA, Landauer, McNamara, Dennis, & Kintsch, 2007), the Topic Model (Griffiths & Steyvers, 2002, Blei, Ng, & Jordan, 2003), Sparse Nonnegative Matrix Factorization (Xu, Liu, & Gong, 2003) and the Constructed Semantics Model (Kwantes, 2005).

Word Overlap: Simple word overlap was used as a baseline in this research. It is the only model that does not use a corpus or knowledge base. Instead, it is a word co-occurrence model. Term frequencies are calculated for each paragraph in the dataset, and similarities are then measured as cosines (see Equation 1) of the resulting paragraph vectors.

$$\cos\theta = \frac{\mathbf{v_1} \cdot \mathbf{v_2}}{\|\mathbf{v_1}\| \, \|\mathbf{v_2}\|} \tag{1}$$

The Vectorspace model (Salton, Wong & Yang, 1975): The Vectorspace model assumes that terms can be represented by the set of documents in which they appear. Two terms will be similar to the extent that their document sets overlap. To construct a representation of a document, the vectors corresponding to the unique terms are multiplied by the log of their frequency within the document, and divided by their entropy across documents, and then added. Using the log of the term frequency ensures that words that occur more often in the document have higher weight, but that document vectors are not dominated by words that appear very frequently. Dividing by the entropy or inverse document frequency (IDF) reduces the impact of high frequency words that appear in

many documents in a corpus. Similarities are measured as the cosines between the resultant vectors for two documents.

Latent Semantic Analysis (LSA, Landauer, McNamara, Dennis, & Kintsch, 2007): LSA starts with the same representation as the Vectorspace model - a term by document matrix with log entropy weighting[5]. In order to reduce the contribution of noise to similarity ratings, however, the raw matrix is subjected to singular value decomposition (SVD). SVD decomposes the original matrix into a term by factor matrix, a diagonal matrix of singular values, and a factor by document matrix. Typically, only a small number of factors (e.g., 300) are retained. To derive a vector representation of a novel document, term vectors are weighted, multiplied by the square root of the singular value vector and then added. As with the Vectorspace model, the cosine is used to determine similarity.

The Topic Model (Topics, Griffiths & Steyvers, 2002, Blei, Ng, & Jordan, 2003): The Topic Model is a Bayesian approach to document similarity that assumes a generative model in which a document is represented as a multinomial distribution of latent topics, and topics are represented as multinomial distributions of words. In both cases, Dirichlet priors are assumed. The parameters of these models can be inferred from a corpus using either Markov Chain Monte Carlo techniques (MCMC, Griffiths & Steyvers, 2002) or variational Expectation Maximization (Blei, Ng, & Jordan, 2003). We implemented the former. Ideally, document representations should then be calculated by running the MCMC sampler over a corpus augmented with information from the new document. To do this on a document by document basis is impractical. In the first instance, we choose to run the sampler over the corpus and then average the word distributions to calculate topic distributions for novel documents. Later in the paper, we investigate the impact of this decision by running the sampler over an augmented corpus containing all of the dataset

paragraphs.

To calculate the similarity of the topic distributions representing documents, we employed both the Dot Product (see Equation 2) and Jensen-Shannon Divergence (JSD, see Equation 3). While the Dot Product was employed for convenience, the JSD is a symmetric form of the Kullback-Leibler Divergence (D) which derives from information theory and provides a well motivated way of comparing probability distributions.

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i \tag{2}$$

$$JSD(P \parallel Q) = \frac{1}{2} D(P \parallel M) + \frac{1}{2} D(Q \parallel M) \tag{3}$$
$$where \quad M = \frac{1}{2}(P + Q)$$
$$and \quad D(P \| Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Sparse Nonnegative Matrix Factorization (SpNMF, Xu, Liu, & Gong, 2003): Nonnegative Matrix Factorization is a technique similar to LSA, which in this context creates a matrix factorization of the weighted term by document matrix. This factorization involves just two matrices - a term by factor matrix and a factor by term matrix - and is constrained to contain only nonnegative values. While nonnegative matrix factorization has been shown to create meaningful word representations using small document sets, in order to make it possible to apply it to large collections we implemented the sparse tensor method proposed by Shashua and Hazan (2005). As in LSA, log entropy weight term vectors were added to generate novel document vectors and the cosine was used as a measure of similarity.

The Constructed Semantics Model (CSM, Kwantes, 2005): The final model considered was the constructed semantics model (Kwantes, 2005). CSM was developed as

a global-matching model of semantic memory derived from the MINERVA 2 architecture of episodic memory. Therefore, CSM is unique in that it was created primarily as a cognitive model to explain the emergence of semantics from experience. To this end, CSM uses a retrieval operation on the contexts in which words occur to generate semantic representations. It operates by taking the term by document matrix (using just log weighting) and multiplying it by its transpose. Consequently, terms do not have to appear together in order to be similar as is the case in the Vectorspace model. Again terms are added to create novel document vectors and the cosine is used as a measure of similarity.

### *2.2. The Datasets*

Two datasets of human ratings of paragraph similarity were used in this study. The first, which we will refer to as the WENN dataset was composed of similarity ratings generated by subjects comparing celebrity gossip paragraphs taken from the World Entertainment News Network. The second dataset, which we will refer to as the Lee dataset, was archival data collected by Lee et al. (2005).

*2.2.1. The WENN dataset*. Students who were recruited by advertising the experiment on a local university campus, along with 17 employees of Defence Research and Development Canada - Toronto (DRDC), provided paragraph similarity ratings to form the WENN dataset. Participants were paid CA$16.69 for taking part in the study. Twenty-three[6] single paragraphs were compared by participants that were selected from the archives of World Entertainment News Network (WENN) made available through the Internet Movie Database[7] (see Appendix A in the Supplementary Material file). Paragraphs were not chosen randomly. First, each paragraph was chosen to be approximately 100 words long. The median number of words contained in paragraphs in

the WENN dataset was 126, paragraph lengths ranging from 79 to 205 words. Paragraphs were also chosen in such a way to ensure that at least some of the paragraphs possessed topical overlap. For example, there was more than one paragraph about health issues, drug problems, stalkers, and divorce among those represented in the stimuli.

Participants were shown pairs of paragraphs, side by side, on a personal computer monitor. Pairs were presented one at a time. For each pair, participants were asked to rate, on a scale of 0 to 100, how similar they felt the paragraphs were to each other. Participants were not given any instructions as to the strategy they should use to make their judgments. Once a similarity judgment had been made, the next pair was presented. Each participant rated the similarity of every possible pairing of different paragraphs for a total of 253 judgments. Pearson correlations were calculated between participants' pairwise comparisons of the paragraphs in the WENN dataset, the average of these correlation coefficients (0.466) indicates that there was only moderate inter-rater reliability for the WENN dataset.

*2.2.2. The Lee dataset*. Lee et al. (2005) recorded observations of paragraph similarity made by 83 Adelaide University students to form the Lee dataset. The dataset consists of ten independent ratings of the similarity of every pair of 50 paragraphs selected from the Australian Broadcasting Corporations news mail service (see Appendix B in the Supplementary Material file), which provides text e-mails of headline stories. The 50 paragraphs in the Lee dataset range in length from 56 to 126 words, with a median of 78.5 words. Pairs of paragraphs were presented to participants on a computerized display. The paragraphs in the Lee dataset focused on Australian and international "current affairs", covering topics such as politics, business, and social issues. Human ratings were made on a 1 (least similar) to 5 (most similar) scale. As mentioned above, Lee et al. (2005)

calculated an inter-rater reliability of 0.605.

*2.3. Domain-chosen corpora: WENN (2000-2006) & Toronto Star (2005)*

Two corpora were chosen to act as knowledge bases for the semantic models to allow similarity estimates to be made on the paragraphs contained in the WENN and Lee datasets. The larger set of 12787 documents collected from WENN between April 2000 and January 2006 was considered a relevant backgrounding corpus for the 23 paragraphs contained in the WENN dataset, this larger set of documents is henceforth called the WENN corpus. It was not possible to resource the original set of 364 headlines and précis gathered by Lee et al. (2005) from the ABC online news mail service. Therefore, in an attempt to provide a news media-based corpus that was similar in style to the original corpus of ABC documents used by Lee and colleagues, articles from Canada's Toronto Star newspaper were used. Moreover, the Toronto Star corpus comprised of 55021 current affairs articles published during 2005.

Initially, both corpora were preprocessed using standard methods: characters converted to lower case, numbers were zeroed (i.e., 31 Jan 2007 became 00 jan 0000), punctuation and words from a standard stop-list (see Appendix C in the Supplementary Material file) were removed, and words that appear only once in a corpus were also removed. Descriptive statistics for both the WENN corpus and the Toronto Star corpus are displayed in Appendix D (see the Supplementary Material file).

## 3. Study One. Comparison of models on domain-chosen corpora

Comparisons made between all semantic models and human evaluations of paragraph similarity for both datasets are presented in the following two subsections of this paper. For the more complex models (LSA, Topics and SpNMF) one must select a

number of dimensions in which to calculate similarities. Performance is likely to be influenced by this choice, therefore in each case comparisons were made using 50, 100 and 300 dimensional models.

*3.1. WENN dataset & WENN Corpus*

Using the WENN corpus, correlations between similarity ratings made by humans and the models on paragraphs in the WENN dataset were low (see Fig. 1) for all models except the simple word overlap (0.43). Of the other models, CSM (0.26) and LSA at 50 dimensions (0.21) performed best. Using the Jensen-Shannon metric improved the performance of the Topic Model in all cases when compared to the dot product measure of similarity. It could be argued that both Vectorspace ($r = 0.17$, $t_{(250)} = 1.61$, n.s.)[8] and LSA at 50 dimensions ($r = 0.21$, $t_{(250)} = 1.05$, n.s.) performed as well as the CSM on this document set. For LSA, the Topic Model and SpNMF, increasing the dimensionality or number of topics did not significantly increase or decrease model performance at this task (see Table E1 in the Supplementary Material file).

Insert Fig. 1 about here

*3.2. Lee dataset & Toronto Star Corpus*

Again except for the word overlap (0.48), the correlations between similarity ratings made by human participants and the models on the paragraphs in the Lee dataset were very low (see Fig. 2). CSM and SpNMF (300 dimensions) were the next best performing models, correlating 0.15 and 0.14 with human judgments, respectively. Also, Vectorspace had higher correlations than both LSA and the Topic Model using the dot product

similarity measure. In 9 out of 12 possible comparisons, increased dimensionality produced significantly better estimates of paragraph similarity by models when compared to human ratings (see Table E2 in the Supplementary Material file).

---

Insert Fig. 2 about here

---

*3.3. Summary of Study One*

Overall, the simple word overlap model outperformed the more complex semantic models when paragraph similarities were compared to human judgments made on both WENN and Lee datasets. On the Lee dataset, semantic models generally performed better when semantic spaces were compiled with higher dimensionality. However, when model dimensionality was increased on the WENN dataset, a similar increase in performance was not found. The generally poor results for the more complex models could be the product of at least one of the following circumstances:

a) the models are unable to generate similarity calculations which are comparable with human judgments.

b) the preprocessing of corpora may have been inadequate, to the extent that noise remained in the corpora which prevented the semantic models from making reasonable estimates of paragraph similarity.

c) or, the corpora did not represent the knowledge required to make similarity estimates on the paragraph contained in WENN and Lee document sets.

Other studies have reported more encouraging results when comparing estimates of paragraph similarity generated by semantic models and humans (Lee, Pincombe & Welsh, 2005 and Gabrilovich & Markovitch, 2007). Therefore, the first possible conclusion is

likely to be inaccurate, indicating semantic models can make a reasonable estimate of the similarity of paragraphs when compared to human judgments. While this was not the case in this study, poor performance by the semantic models may have been driven by a sub-optimal match between the background corpus and the paragraphs being tested. The likelihood of this scenario is supported by the generally low correlations with human results obtained by all of the models that required a background corpus. The following three studies explore the latter two possibilities. In Study Two, a more stringent corpus preprocessing method is used to improve on the results presented in Study One. In Study Three, Wikipedia is used to generate better backgrounding corpora, and this method again improves model estimates of paragraph similarity when compared to the human judgments. Then, in Study Four, paragraphs from the datasets are added to the models' knowledge base to again improve model performance at this task.

## 4. Study Two: Corpus Preprocessing

Generally, corpus preprocessing identifies words that are likely to be informative to the semantic model. In the field of information retrieval there have been many types of sophisticated term selection functions employed by researchers (Sebastiani, 2002, p. 15). Other methods such as employing a stop-list are less complex, requiring no mathematical calculation, and simply remove words from the corpus which are deemed uninformative by the researcher. Stop-lists are usually applied to remove words such as articles, pronouns and conjunctions (Moed, Glänzel, & Schmoch, 2004). Bullinaria and Levy (2006) found that stop-lists reduced model performance when the textual unit under comparison is at a word-word level (such as the TOEFL task described above). However, working with paragraph comparisons, Pincombe (2004) states that "[u]se of a stop word

list almost always improved performance" when comparing models estimates of similarity and human judgments (p. 1). A closer inspection of the stop-list (Appendix F in the Supplementary Material file) and preprocessing techniques (p. 14) used by Pincombe (2004)[9] was conducted. This review revealed that single letters had been removed by the author and only alphabetical characters had been used in his corpora. The difference between the preprocessing used in Study One (allowing the inclusion of zeroed numbers and single characters) and that used in Pincombe's research begs the question:

Can the removal of single letters and numbers from the background corpus improve a semantic model's ability to estimate paragraph similarity?

It is possible that the presence of these types of information (numbers and single letters) in a corpus can create noise for the models. For example, the American Declaration of Independence in 1776 has little to do with Elvis Presley's birthday in 1935. Although using the preprocessing method of zeroing numbers, models comparing texts that describe these two occasions would erroneously find some similarity between them. Moreover, the zeroing of the aforementioned dates could also suggest commonality with a document describing the distance between two cities, obviously creating noise in the corpus even if this new document described a 1000 mile drive between Philadelphia (Pennsylvania) and Tupelo (Mississippi). Similarly, the 'Js' in 'J F K' and 'J K Rowling' should not indicate semantic similarity between documents that make reference to these well known individuals. Therefore, the removal of these items may benefit a model's ability to perform similarity ratings between paragraphs.

### *4.1. Removing numbers & single letters*

All numbers and single letters were removed from both the WENN and Toronto Star corpora[10] to test the hypothesis that removing these characters would improve the

semantic models' performance when similarity ratings were compared to human judgments. Fig. 3 and Fig. 4 display comparisons between the results generated in Study One (ALL) and the results for spaces compiled on corpora without number and single letters (NN-NSL, No numbers - No Single Letters). Only the results for models compiled at 300 dimensions (where dimensionality is a parameter of the model) are displayed in these figures. It should be noted, while the models compiled at 300 dimensions generally[11] produced the best results, models compiled at both 50 and 100 dimensions displayed an identical trend (see Table G1 in the Supplementary Material file) of better performance when using the more stringent preprocessing method.

Insert Fig. 3 about here

Insert Fig. 4 about here

Although it may seem counter intuitive to remove information from a knowledge base or corpus, the removal of numbers and single letters improved correlations between human judgments and similarity ratings produced from models in nearly all comparisons that were made for both the WENN (see Fig. 3) and Lee (see Fig. 4) datasets. The only model that did not improve in performance was CSM on the WENN dataset. This difference for CSM between ALL (0.26) and NN-NSL (0.16) corpora was significant ($t_{(250)} = -2.48$, $p < 0.05$). A more promising trend was displayed by the other models, especially on the WENN dataset with the LSA (0.48) and SpNMF (0.43) models performing best of the more complex semantic models. However, this trend was also

displayed by the simple word overlap model which continued to clearly outperform the other models. When numbers and single letters were removed from the paragraphs used by the overlap model, correlations between this model and the human judgments improved to 0.62 on the WENN dataset and 0.53 on the Lee dataset. In 4 out of 12 comparisons on the WENN dataset, and 5 out of 12 comparisons on the Lee dataset, increased dimensionality led to significant improvements to models' performance (see Tables G1 and G2 in the Supplementary Material file).

Notwithstanding this general improvement in the more complex semantic models' performance, correlations with human judgments of similarity were still low using the Toronto Star (NN-NSL) corpus on the Lee dataset, with the highest being the Vectorspace model (0.2). This suggests that while corpus preprocessing was hindering the models' ability to provide reasonable estimates of paragraph similarity, there are also other factors that are impeding the models' performance. Clearly, the information and themes contained within corpora certainly constrain the performance of semantic models. However, suitable knowledge bases are not always easy to obtain. In an attempt to address this issue, the third study examines an alternative method of generating corpora that draws sets of knowledge-domain related documents (sub-corpora) from the online encyclopedia Wikipedia.

## 5. Study Three: A better knowledge base?

Smaller, more topic focused, sub-corpora may provide context for polysemous words, that may otherwise take on several meanings in a larger corpus. To this end, Wikipedia was utilized as a generic set of documents from which smaller targeted sub-corpora could be sampled and compiled. Wikipedia is maintained by the general

public, and has become the largest and most frequently revised or updated encyclopedia in the world. Critics have questioned the accuracy of the articles contained in Wikipedia, but research conducted by Giles (2005) did not find significant differences in the accuracy of science-based articles contained in Wikipedia when they were compared to similar articles contained in the Encyclopedia Britannica. Furthermore, the entire collection of Wikipedia articles are available to the general public and can be freely downloaded[12]. All Wikipedia entries current to March 2007 were downloaded for this research. In total there were 2.8 million Wikipedia entries collected; however, the total number of documents was reduced to 1.57 million after the removal of incomplete articles contained in the original corpus. Moreover, incomplete articles were identified and removed if they contained phrases like "help wikipedia expanding" or "incomplete stub". The resulting Wikipedia corpus was further preprocessed in the same manner as the NN-NSL corpora in Study Two: removing stop-words, punctuation, words that only appeared once in the corpus, and finally removing all numbers and single letters.

To enable the creation of sub-corpora, Lucene[13] (a high performance text search engine) was used to index each document in the Wikipedia corpus. Lucene allows the user to retrieve documents based on customized queries. Like the search results provided by Google, the documents returned by Lucene are ordered by relevance to a query.

Targeted queries were created for each paragraph rated by humans in the WENN dataset. This WENN-based query was constructed by removing stop-words and punctuation from the title[14] that accompanied each paragraph, and then joining the remaining words with "OR" statements (see Appendix H in the Supplementary Material file). In contrast, the query devised for the paragraphs in the Lee dataset was more complex. For the Lee-based query, the researcher chose several descriptive keywords[15]

for each paragraph in the Lee dataset, and used “AND” and ‘OR“ operators to combine these keywords. Moreover, the Lee-based query used Lucene’s ‘star’ wild-card operator to return multiple results from word stems. For example, the stem and wild-card combination “research*” would match documents containing the words “research”, “researcher”, and “researchers” (see Appendix I in the Supplementary Material file).

*5.1. Wikipedia Sub-corpora*

Four sub-corpora were created using the Lucene queries (described above) on the Wikipedia document set. For each dataset (WENN & Lee), a 1000 document and a 10000 document sub-corpus was generated. The structure of the Wikipedia articles contained in these sub-corpora was substantially different from the documents contained in either the WENN or Toronto Star corpora (see Table D1 in the Supplementary Materials file). Wikipedia articles tend to be longer in format, with documents that approximate the length of a short essay (on average 1813 to 2698 words per document). In contrast, the documents contained in both the WENN and Toronto Star corpora are similar in length to a journal article’s abstract (on average 74 to 255 words per document). In addition to the Wikipedia documents being generally much longer than the WENN or Toronto Star documents, the Wikipedia documents also contain on average many more unique words.

The greater size and complexity of the Wikipedia documents may produce noise for the semantic models. However, Lee and Corlett’s (2003) findings indicate that decisions about a document’s content can be made using only the beginning of a document’s text. In their study of Reuters’ documents, words found in the first 10 percent of a document’s text were judged to hold greater ‘mean absolute evidence’ characterizing a document’s content. Lee and Corlett calculated the ‘evidence’ value of a word given a particular topic. This calculation was made by comparing how often a word appeared in documents related

to a topic, relative to the word's frequency in documents that were not topic-related. Their finding may reflect a generally good writing style found in Reuters' documents, where articles may begin with a précis or summary of the information that is contained in the rest of the document. Documents in a web-based medium such as Wikipedia, may also conform to this generalization. Intuitively, it seems likely that important descriptive information displayed on a web page will be positioned nearer the top of a page (probably within the first 300 words), so as not to be over-looked by the reader as the web page scrolls or extends beneath screen[16].

To explore the possible effect of document length (number of words) on semantic models, corpora were constructed that contained the first 100, 200, 300 and all words from the Wikipedia sub-corpora. To illustrate this point, if the preceding paragraph was considered a document, in the first 100 word condition this document would be truncated at "...by comparing how often a word appeared in". Furthermore, to test if corpus size influenced the similarity estimates generated by the semantic models, performance was compared on the 1000 and 10000 sub-corpora for both datasets. Thus, making a 2 x 4 design (number of documents in a corpus BY number of words in each document) for each dataset. Each sub-corpus was compiled using LSA at 300 dimensions. LSA was chosen for its quick compilation speeds and because of the generally good match that has been reported between LSA and human performance on tasks comparing paragraph similarity (Lee et al., 2005; Landauer & Dumais, 1997). Moreover, in general LSA was one of the best performing models that incorporates a knowledge base in the previous studies presented in this paper[17]. This choice of dimensionality is supported by the findings of the first two studies in this paper, where increased dimensionality improved performance.

Insert Fig. 5 about here

Insert Fig. 6 about here

*5.1.1. Document Length*. In general, LSA's performance was better as document length was shortened, with the best results produced by truncating documents length at 100 words. On both datasets, LSA produced the highest correlations with the human similarity judgments using the 1000 document sub-corpora truncated at 100 words (see Fig. 5 and Fig. 6). This configuration produced a result (0.51) that was significantly higher than all other document number and document length combinations for the Lee dataset. On the WENN dataset, the correlation for the 1000 document corpora with documents truncated at 100 words was higher than all other cases; however, this result was not significantly higher in several cases. On both datasets, truncating documents at 100 words produced significantly higher correlations than the ALL word conditions (where document length was not truncated). These results show that improvements to model performance can be achieved by truncating documents to 100 words, and this improvement supports the earlier findings of Lee & Corlett (2003).

*5.1.2. Number of Documents*. LSA's performance on both datasets was best using the smaller 1000 document sub-corpora. On the Lee dataset, when documents are truncated at 100 words, the performance of LSA is better using the 1000 document sub-corpora than the 10000 document sub-corpora ($t_{(1222)} = 4.44$, $p < 0.05$). On the WENN dataset, when documents are truncated at 100 words performance was also better

for the 1000 document sub-copora, although this difference failed to reach significance ($t_{(250)} = 1.63$, n.s.).

*5.2. All models compared on Wikipedia sub-corpora*

The results presented in Study Two of this paper for models using the WENN (NN-NSL) and Toronto Star (NN-NSL) corpora have also been included in the findings presented in Fig. 7 and Fig. 8 as points of comparison to judge the effectiveness of creating the 1000 and 10000 document sub-corpora from Wikipedia.

When the results for both the WENN and Lee datasets are taken into consideration, again none of the more complex semantic models performed significantly better than the simple word overlap model. While the best performing model on the Lee dataset was Vectorspace (0.56) using the Wikipedia 10000 document corpus, this was not significantly different ($t_{(1222)} = 1.31$, n.s.) from the word overlap model's correlation (0.53) with human judgments. As is displayed in Fig. 7 and Fig. 8, of the corpus-based models Vectorspace, LSA and SpNMF performed the best on both datasets. It is unclear whether using the Jensen-Shannon metric as opposed to dot product measure with the Topic Model produced better results. On the Lee dataset, Topic Model with dot product (0.48) using the 1000 document Wikipedia corpus significantly outperformed Topics model with the Jensen-Shannon metric (0.42) using the 10000 document Wikipedia corpus ($t_{(1222)} = -2.08$, $p < 0.05$). However, using the WENN (NN-NSL) corpus, there was not a significant difference between the two Topic Model similarity measures ($t_{(250)} = 0.53$, n.s.) on the WENN dataset.

---

Insert Fig. 7 about here

---

Insert Fig. 8 about here

LSA performed well using both the WENN (NN-NSL) and Wikipedia-based Lee corpora. Given that LSA is built on Vectorspace, it is encouraging to see that in the case of the WENN dataset dimensionality reduction improved this LSA's performance (0.48) when compared to Vectorspace (0.41). However, this improvement was not found consistently, as indicated by the higher correlation with human judgments on Lee dataset achieved by Vectorspace using either 1000 and 10000 document Wikipedia-based corpora (Fig. 8).

Using the WENN (NN-NSL) corpus as a knowledge base allowed the semantic models to produce better estimates of human similarity judgments than could be obtained using either 1000 or 10000 document Wikipedia-based corpora on the WENN dataset. In contrast, corpora retrieved from Wikipedia allowed the models to perform much better when making estimates of paragraph similarity on the Lee document set (see Fig. 8). For corpus-based models, the 10000 document Wikipedia corpus was found to produce the highest correlation with human ratings on the Lee document set (Vectorspace 0.56), however in the majority of cases the 1000 document Wikipedia corpora was associated with better model performance at this task. All results presented thus far have consistently shown that the Toronto Star has provided a poor knowledge base on which to assess the paragraphs contained in Lee dataset. These results indicate that when domain-chosen corpora are not a good fit to the knowledge required to make accurate estimates of similarity on paragraphs, using corpora drawn from Wikipedia can improve model performance.

## 6. Study Four: Corpora that include the dataset paragraphs

In the empirical studies we have reported, subjects were presented with document pairs to be rated. Documents were repeated in different pairs, so for the majority of ratings subjects had already been exposed to all of the test documents. In the previous studies, paragraphs contained in the WENN dataset were included in the WENN corpora, but not for the corpora used by models on the Lee dataset. Consequently, the models were at a disadvantage relative to participants. This inclusion of dataset paragraphs is potentially important for models like the Topic Model where context can select for the appropriate meaning of a word. To evaluate the efficacy of including stimulus paragraphs into the semantic models' knowledge base as a method of corpus improvement, the 50 Lee dataset paragraphs were added to the most effective corpora with the most effective preprocessing found in the previous studies for the Lee dataset.

For this study, the 50 Lee paragraphs were prepended to both the 1000 and 10000 document Wikipedia corpora. These revised corpora were preprocessed using the same techniques described in Study Three for the Wikipedia sub-corpora. While the 50 Lee paragraphs were not truncated at 100 words, preprocessing was used to remove punctuation, stop-list words, words that only appear once on the document set, numbers and single letters. After preprocessing, the smaller corpus contained 1,050 documents with 8,674 unique words and 100,107 tokens, and the larger corpus held 10,050 documents comprised of 37,989 unique words from a total of 942,696 tokens.

---

Insert Fig. 9 about here

---

Adding the 50 Lee paragraphs to the Wikipedia 1000 corpora significantly improved

correlations between model estimates and human judgments of similarity in nearly all cases (see Table K1 in the Supplementary Material file). While the Topics model improved one point, using the dot product measure, there was not a significant improvement using the Jensen-Shannon metric. The greatest improvement in model performance was displayed by Vectorspace which increased from 0.55 to 0.67, and LSA which rose from 0.51 to 0.60 (see Fig. 9).

Both significant performance increases and decreases were produced for all models by prepending the 50 Lee paragraphs to the 10000 document Wikipedia corpora (see Table K2 in the Supplementary Material file). While all differences were significant when compared to non-augmented Wikipedia sub-corpora, the actual differences in performance were small for most models. In general, these differences ranged from between 0.001 to 0.02 with the exception of Topics model using the Jensen-Shannon metric which went up from 0.42 to 0.49 when the 50 Lee paragraphs were added to the Wikipedia 10000 corpus (see Fig. 10).

---

Insert Fig. 10 about here

---

## 7. Overall Summary

In Study One, moderate correlations were found between the word overlap model (WENN: 0.43, Lee: 0.48) and human judgments of similarity on both datasets. However, weaker performance was displayed by all of the more complex models when similarity estimates were compared on both the WENN (highest CSM, 0.26) and Lee (highest CSM 0.15) datasets. It was postulated that the semantic models' performance may have been constrained by factors such as corpus preprocessing and a poorly represented knowledge

domain (in the case of the Toronto Star corpus and the Lee dataset). In Study Two, the importance of corpus preprocessing was highlighted, removing the numbers and single letters from corpora improved correlations with human judgment on both datasets for all models with the exception of CSM. After the removal of these characters, the best performing of the more complex models were LSA (0.48) on the WENN dataset and Vectorspace (0.20) on the Lee dataset. However, the corpus-based models still failed to outperform the word overlap model, which also improved with the removal of numbers and single letters on both datasets (WENN: 0.62, Lee: 0.53).

In some ways it is unsurprising that the models' performance in this study was better on the WENN dataset than the Lee dataset, because the paragraphs used in similarity judgments were drawn from the greater set of documents contained in the WENN corpus. That is, in the case of the WENN set there was a better match between paragraphs that were compared (WENN dataset) and the models' knowledge base (the WENN Corpus). Conversely, the Toronto Star articles did not provide a good approximation of the knowledge required to make reliable inferences regarding the similarity of paragraphs contained in the Lee dataset. While the Toronto Star corpus contains extracts of current affairs, these articles (published in 2005) must vary substantially from the précis published in 2001 that are contained in the ABC news mail service that was used by Lee et al. (2005).

In an attempt to obtain a better representation of the knowledge base required to make accurate paragraph similarity comparisons, in Study Three Wikipedia sub-corpora were generated to use on each dataset. The Wikipedia documents were found to be much longer and more like short essays than the summary or abstract length documents found in the WENN and Toronto Star corpora. Guided by the research findings of Lee and Corlett

(2003), it was found that Wikipedia documents truncated at 100 words provided better corpora for LSA at 300 dimensions than when using all of the words contained in these documents. LSA's performance was also better using the smaller 1000 document Wikipedia sub-corpora. The decision to use 300 dimensions was in part based on the results of Study One and Study Two, which indicated that increased dimensionality often led to significant performance gains when model estimates of paragraph similarity were compared to human ratings.

Based on these findings, spaces were compiled for the models using Wikipedia corpora that contained documents truncated at 100 words. The semantic models' performance on the WENN dataset did not improve using these Wikipedia sub-corpora when compared to results achieved by models using the WENN corpus. However, there was a substantial improvement by nearly all models (except CSM) when similarity estimates were compared on the Lee dataset. Using the Wikipedia sub-corpora, the best performing of the more complex models on the Lee dataset were Vectorspace using both 1000 documents (0.55) and 10000 documents (0.56) and SpNMF using 1000 documents (0.53); all of which approach the inter-rater reliability (0.6) recorded for Lee and colleagues participants (Lee et al., 2005). The decrement in performance seen using the Wikipedia sub-corpora, when compared to the WENN corpus on the WENN dataset, is again somewhat expected given that the documents in the WENN dataset were selected from the WENN corpus. When the results on both the WENN and Lee datasets are considered, Vectorspace, LSA and SpNMF were the best performing of the corpus-based models. That said, even using corpora that allowed models to perform on a comparable level with the inter-rater reliability found in the WENN dataset, and approaching that calculated for the Lee dataset, these models still could not significantly outperform the

simple word overlap model when estimating the similarity of paragraphs in comparison to human performance at this task.

The final study explored what effect including the dataset paragraphs into a corpus would have on models' performance. This assessment was only undertaken for the Wikipedia corpora used on the Lee dataset, as the WENN documents were already included in the WENN corpora in previous studies. In particular, the Topic Model performance was expected to increase; however, this improvement in performance was only observed for the Topic Model using the Dot Product measure of similarity. Generally, performance increases associated with the inclusion of the 50 Lee paragraphs were greater on the smaller 1050 document Wikipedia corpus when compared to those observed on the 10050 Wikipedia corpus. It is possible that any benefit to a model's performance produced by adding these 50 paragraphs is negated by the volume of terms contained in the larger corpus. Overall, the best performance was observed for Vectorspace (0.67) and LSA (0.60) using the 1050 document Wikipedia corpus containing the 50 Lee paragraphs. It was interesting to note, that LSA's performance using the smaller Wikipedia corpus and 50 Lee paragraphs was almost exactly the same as the inter-rater reliability calculated for the Lee dataset. Furthermore, using this augmented 1050 document Wikipedia corpus, both LSA ($t_{(1222)} = 3.20$, $p < 0.05$) and Vectorspace ($t_{(1222)} = 7.81$, $p < 0.05$) significantly outperformed the overlap model (0.53) when estimates of paragraph similarity were compared to the human judgments contained in the Lee dataset.

Fig. 11 displays scatterplots from the two best performing models on WENN and Lee datasets. It was surprising that on both datasets, the simple word overlap model was among the two best performing models. As is illustrated by Fig. 11B and D, the word overlap model is generally capturing human responses that have been made on paragraphs

which have low or no similarity. It is also interesting to note that on the WENN dataset, LSA using the WENN corpus (NN-NSL) has in all cases estimated some similarity between the paragraph pairs (see Fig. 11A). This may indicate that greater dimensionality is needed by LSA to adequately delineate the themes presented in the WENN corpus documents. At another level, because the WENN paragraphs all focus on "celebrity gossip", to some extent they may all be considered related. Alternatively, on the Lee dataset, Vectorspace appears to have provided a relatively good match to the average human estimates of paragraph similarity (see Fig. 11C).

---

Insert Fig. 11 about here

---

## 8. Discussion

Quite surprisingly, the simplest models (Vectorspace and word overlap) were the best performing models examined in this research, both exceeding the inter-rater reliability calculated for human judgments. While exceeding the inter-rater reliability is an important milestone, it is possible for a model to perform better. The model is compared against the average rating of the subjects, which eliminates a significant amount of variance in the estimates of the paragraph similarities, whereas the inter-rater reliability is the average of the pairwise correlation of the subjects. On the WENN dataset, the overlap model (0.62) exceeded the inter-rater reliability (0.47). Similarly the Vectorspace model (0.67) using a corpus containing both truncated Wikipedia documents and the 50 Lee paragraphs also exceeded the inter-rater reliability found for the Lee dataset (0.605).

The Vectorspace model's performance on the Lee dataset using the smaller Wikipedia corpus that included the 50 Lee paragraphs was particularly encouraging.

While the overlap model's good performance at these tasks can largely be accounted for by its ability to capture human ratings on paragraph pairs with low or no similarity, the Vectorspace model appeared to provide good estimates of both the similarity and dissimilarity of the Lee paragraphs when compared to human ratings. That said, the Vectorspace model did not perform as well on the WENN dataset, when compared to estimates produced by either the overlap model or LSA. However, the finding that no model performed as well as the overlap model on this dataset, might indicate that even though the best results for the WENN dataset were found for most corpus-based models using the WENN corpus (NN-NSL), that this corpus still did not provide an adequate term representation for the models. Furthermore, it is possible that a better match to the background knowledge needed by models for the WENN paragraphs may have been accomplished had a more complex Lucene query been used to retrieve relevant Wikipedia documents.

One possible explanation for the success of the overlap and Vectorspace models in these studies may be found in the framework of the experiments. In each experiment, participants made pairwise comparisons of paragraphs displayed on a computer monitor. The side-by-side positioning of these paragraph pairs may have encouraged keyword-matching (or discrimination) between the paragraphs by the participants. That is, it is possible that the participants were skimming paragraphs for keywords with which they could make similarity judgments. Another related strategy which could result in the similar outcome, would be to read one paragraph thoroughly and then to skim the comparison paragraph for salient words presented in the first text. Masson (1982) indicates that when skimming, readers miss important details in newswire texts, and that visually unique features of text such as place names may increase efficiency of skimming

as a reading strategy. Given that names of people and places were certainly present in all paragraphs presented to participants in this research, commonalities between participants' similarity estimates (and also those of the overlap and Vectorspace models) may also be influenced by these proper nouns. In future research, eye-tracking technology could be employed to elucidate the possibility of skimming strategies in this type of experimental task. Alternatively, paragraphs could be presented in a serial sequence, rather than concurrently, and time spent reading each paragraph might act as an indicator of reading strategy.

In the introduction, we categorized the materials used in this class of research as having four types of textual unit: words, sentences, single paragraphs, and chapters or whole documents. Past research has indicated that the Topic Model performs better at word association tasks than LSA. Moreover, researchers have shown that Topic Model's ability to accommodate homographs is superior to other models at the single word textual unit level (Griffiths et al., 2007). While the ability to discriminate the intended meaning of ambiguous words is certainly desirable, it is possible that this attribute is not a prerequisite for successful model performance with larger textual units such as paragraphs. This may be because textual units such as sentences and paragraphs allow models access to a range of non-ambiguous words whose informativeness may compensate a model's inability to capture the meaning of more ambiguous words (Landauer & Dumais, 1997; Choueka & Lusignan, 1985). In the studies reported above, four models (word overlap, Vectorspace, LSA, and SpNMF) that do not capture this type of word ambiguity all outperformed the Topic Model when compared to human ratings at the task of estimating similarity between paragraphs.

Besides a model's ability to make good approximations at human similarity

judgments, another factor that must be considered when evaluating the usefulness of these semantic models is the ability to produce interpretable dimensions. For example, one of the criticisms of LSA is that the dimensions it creates are not always interpretable (Griffiths et al., 2007). Similarly, word overlap, Vectorspace and CSM do not employ any dimensionality reduction, and thus provide word vectors that are difficult to interpret. In contrast, both SpNMF and Topic Model return interpretable dimensions. To illustrate this point, Table J3 in the Supplementary Material file displays a sample of the dimensions created for the 10000 document Lee-based Wikipedia corpus. As is made clear in this table, it would be easy to meaningfully label any of these dimensions.

Given its generally good approximations of human judgment and ability to provide interpretable dimensions, SpNMF could be regarded as one of the best models examined in this article. However, its slow compilation of spaces would certainly need to be addressed for it to be generally useful in either a research or an applied setting. In comparison to the Vectorspace model which takes 24 seconds to compile a space of the King James Bible using a 2.4 GHz CPU, the SpNMF model is very computationally expensive taking just under 8 hours. Future research may be able to utilize parallel programming techniques[18] to sequence SpNMF calculations over multiple processing units to reduce the time needed to compile SpNMF spaces, and thus make SpNMF a more feasible model to use in tasks of this kind.

In several ways, CSM was the worst performing model employed in this research. All models performed better than CSM when using either the Wikipedia sub-corpora or WENN corpus (NN-NSL). Also, the matrices contained within CSM spaces can be over two orders of magnitude larger than those compiled by other models. For example, the space compiled by CSM for the 10000 document Wikipedia-based corpus with documents

truncated at 100 words for the Lee dataset was 12Gb in size. In stark contrast, the same corpus compiled by Vectorspace used 84Mb of disk space. While files of this size are not unusable, accessing the dense vectors contained in CSM spaces is slower than retrieving vectors for comparisons using the other models.

One of the key strengths of the simple overlap model that performed so well in this research, is that it is not reliant on a knowledge base, only extracting information from the surface structure of the textual stimuli. This paper has provided examples of the difficulties researchers face when attempting to create a suitable knowledge base for semantic models. This is not to mention the labor intensive process undertaken to collect and format a large corpus. Furthermore, the simple overlap model is not without theoretical underpinning. Max Louwerse, in this issue of TopiCS, has suggested that "support for language comprehension and language production is vested neither in the brain of the language user, its computational processes, nor embodied representations, but outside the brain, in language itself". In arguing his claim, Louwerse provides examples of how first-order co-occurrences of terms can produce similar results to LSA on tasks of categorization. Similarly, it could certainly be argued that to some extent the good performance of the overlap model in the studies presented in this paper support Louwerse's argument.

Overall, dimensionality reduction did not appear to improve the models' estimate of paragraph similarity when compared to results produced by Vectorspace and overlap models. However, LSA's consistent performance, mimicking of human inter-rater reliability, and better performance on the WENN dataset when compared to Vectorspace, all indicate that there is further research that must be done in this area. One aspect of this research that we intend to explore more fully, is the possibility that subsets of participants'

judgment variance can be accommodated by different models. For example, it is possible that participants who are skimming the paragraphs may produce results more similar to either Vectorspace or the overlap model. In contrast, other participants who are reading carefully and not skimming over the text, may produce results that are more similar to those calculated with LSA. While it is not possible to draw these conclusions with any certainty based on our current datasets, eye-tracking technology will be employed in future research to explore these possibilities.

The findings presented in this paper indicate that corpus preprocessing, document length and content are all important factors that determine a semantic model's ability to estimate human similarity judgments on paragraphs. The online, community driven Wikipedia encyclopedia also proved to be a valuable resource from which corpora could be derived when a more suitable domain-chosen corpus is not available. In many applications the hand construction of corpora for a particular domain is not feasible, and so the ability to show a good match between human similarity judgments and machine evaluations is a result of applied significance.

## Acknowledgments

The research reported in this article was supported by Defence Research & Development Canada (grant number – W7711-067985). Also, we would like to thank Michael Lee and his colleagues for access to their paragraph similarity data. Finally, we wish to thank the reviewers, whose helpful comments have greatly improved this paper.

## References

Bell, N., & Garland, M. (2008). *Efficient sparse matrix-vector multiplication on CUDA* (NVIDIA Technical Report No. NVR-2008-004). NVIDIA Corporation. Available at: http://www.nvidia.com/object/nvidia_research_pub_001.html Accessed April 10, 2009.

Berry, M. W., & Browne, M. (2005). Email surveillance using non-negative matrix factorization. *Computational & Mathematical Organization Theory*, *11* (3), 249–264.

Blei, D., Ng, A. Y., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3* (4-5), 993–1022.

Bullinaria, J. A., & Levy, J. P. (2006). Extracting semantic representations from word co-occurrence statistics: a computational study. *Proceedings of the National Academy of Sciences*, *39* (3), 510–526.

Choueka, Y., & Lusignan, S. (1985). Disambiguation by short contexts. *Computers and the Humanities*, *19* (3), 147–157.

Dennis, S. (2005). A memory-based theory of verbal cognition. *Cognitive Science*, *29* (2), 145–193.

Foltz, P. W. (2007). Discourse coherence and LSA. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 167–184). Mahwah, NJ: Lawrence Erlbaum Associates.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, *1* (2). Available at: http://imej.wfu.edu/articles/1999/2/04/index.asp Accessed April 2, 2008.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis. In M. M. Veloso (Ed.), *Proceedings of the 20th international joint conference on artificial intelligence* (pp. 1606–1611). Menlo Park, CA: AAAI Press.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature, 438* (7070), 900–901.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th annual conference of the cognitive society* (pp. 381–386). Mahwah, NJ: Lawrence Erlbaum Associates.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101 (Suppl. 1)*, 5228–5235.

Griffiths, T. L., Tenenbaum, J. B., & Steyvers, M. (2007). Topics in semantic representation. *Psychological Review*, *114* (2), 211–244.

Kintsch, W. (2001). Predication. *Cognitive Science*, *25* (2), 173–202.

Kireyev, K. (2008). Beyond words: Semantic representation of text in distributional models of language. In M. Baroni, S. Evert, & A. Lenci (Eds.), *Proceedings of the ESSLLI workshop on distributional lexical semantics: Bridging the gap between semantic theory and computational simulations* (pp. 25–33). Hamburg, Germany: ESSLLI.

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin and Review*, *12* (4), 703–710.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent

Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104* (2), 211–240.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lee, M. D., & Corlett, E. Y. (2003). Sequential sampling models of human text classification. *Cognitive Science, 27* (2), 159–193.

Lee, M. D., Pincombe, B. M., & Welsh, M. B. (2005). An empirical evaluation on models of text document similarity. In B. G. Bara, L. W. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the cognitive society* (pp. 1254–1259). Mahwah, NJ: Lawrence Erlbaum Associates.

Martin, D. I., & Berry, M. W. (2007). Mathematical foundations behind Latent Semantic Analysis. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 35–55). Mahwah, NJ: Lawrence Erlbaum Associates.

Martin, M. J., & Foltz, P. W. (2004). Automated team discourse annotation and performance prediction using LSA. In S. T. Dumais, D. Marcu, & S. Roukos (Eds.), *HLT-NAACL 2004: Short papers* (pp. 97–100). Boston, Massachusetts, USA: Association for Computational Linguistics.

Masson, M. E. J. (1982). Cognitive processes in skimming stories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8* (5), 400–417.

McNamara, D. S., Cai, Z., & Louwerse, M. M. (2007). Optimizing LSA measures of cohesion. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.),

*Handbook of Latent Semantic Analysis* (pp. 379–400). Mahwah, NJ: Lawrence Erlbaum Associates.

Moed, H. F., Glänzel, W., & Schmoch, U. (2004). *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems*. Secaucus, NJ, USA: Springer–Verlag New York.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The university of South Florida word association, rhyme, and word fragment norms.* Available at: http://w3.usf.edu/FreeAssociation/ Accessed February 2, 2009.

Pincombe, B. M. (2004). *Comparison of human and LSA judgements of pairwise document similarities for a news corpus* (Tech. Rep. No. DSTO-RR-0278). Adelaide, Australia: Australian Defense Science and Techology Organisation (DSTO), Intelligence, Surveillance and Reconnaissance Division. Available at: http://hdl.handle.net/1947/3334 Accessed April 15, 2008.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18* (11), 613–620.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34* (1), 1–47.

Shashua, A., & Hazan, T. (2005). Non-negative tensor factorization with applications to statistics and computer vision. In L. De Raedt & S. Wrobel (Eds.), *Proceedings of the 22nd international conference on machine learning* (pp. 792–799). New York, NY: ACM Press.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87* (2), 245–251.

Toffler, A. (1973). *Future shock.* London, UK: Pan.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In J. Callan, D. Hawking, A. Smeaton, & C. Clarke (Eds.), *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '03)* (pp. 267–273). New York, NY: ACM Press.

Zelikovitz, S., & Kogan, M. (2006). Using web searches on important words to create background sets for LSI classification. In G. Sutcliffe & R. Goebel (Eds.), *Proceedings of the 19th international FLAIRS conference* (pp. 598–603). Menlo Park, CA: AAAI Press.

## Footnotes

[1]A standard stop-list was applied, and only words appearing 10 times or more were included in the final corpus.

[2]http://www.ets.org/

[3]Dividing by the entropy reduces the impact of high frequency words that appear in many documents in a corpus.

[4]http://en.wikipedia.org/

[5]The reader is directed to Martin and Berry (2007) for an example of how to create a term by document matrix for both the Vectorspace model and LSA.

[6]Participants actually compared 25 paragraphs, however a technical fault made the human comparisons of two paragraphs to the rest of the paragraphs in the set unusable.

[7]http://www.imdb.com

[8]Two-tailed significance tests (alpha = 0.05) between non-independent correlations were performed with Williams' formula (T2) that is recommend by Steiger (1980).

[9]These techniques were also used in the Lee et al. (2005) study.

[10]Both corpora had already been preprocessed with standard methods: removing stop-words, punctuation, and words that appear in only one document were also removed.

[11]With the exception of Topic Model using the Jensen-Shannon metric, all models that incorporate dimensionality reduction performed better at 300 dimensions. Topics-JS at 100 topics was 0.29 compared to 0.28 with 300 topics.

[12]http://download.wikimedia.org/enwiki/latest/

[13]PyLucene, is a Python extension that allows access to the Java version of Lucene: http://pylucene.osafoundation.org/

[14]These titles were not included with the WENN paragraphs when similarity

comparisons were made by either humans or the semantic models.

[15]On average, four keywords were chosen per paragraph to form the Lee-based query.

[16]In web usability research and broad-sheet newspaper media terms this positioning is often referred to as being “above the fold”.

[17]In Study Two, LSA similarity estimates correlated 0.48 with human judgments of similarity on WENN document set.

[18]CUDA, the nVidia graphics processing unit technology, presents as an architecture on which these parallel processing gains might be achieved whilst efficiently using sparse matrices (Bell & Garland, 2008).

## Figure Captions

Fig. 1. Correlations (*r*) between the similarity ratings made on paragraphs in the WENN dataset by human raters and the those made by word overlap, LSA, Topics, Topics-JS (with Jensen-Shannon), SpNMF, Vectorspace, and CSM. All models, except word overlap used the WENN corpus. The effects of dimensionality reduction are displayed at 50, 100 and 300 dimensions for the more complex models that incorporate this reductive process. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

Fig. 2. Correlations (*r*) between the similarity ratings made on paragraphs in the Lee dataset by human raters and the those made by word overlap, LSA, Topics, Topics-JS (with Jensen-Shannon), SpNMF, Vectorspace, and CSM. All models, except word overlap used the Toronto Star corpus. The effects of dimensionality reduction are displayed at 50, 100 and 300 dimensions for the more complex models that incorporate this reductive process. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

Fig. 3. Correlations between similarity estimates made by human and models on paragraphs in the WENN dataset. Models that employ a knowledge base used the WENN corpus. “ALL” depicts standard corpus preprocessing used in Study One, “NN-NSL” corpora have also had numbers and single letters removed. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

Fig. 4. Correlations between similarity estimates made by human and models on

paragraphs in the Lee dataset. Models that employ a knowledge base used the Toronto Star corpus. “ALL” depicts standard corpus preprocessing used in Study One, “NN-NSL” corpora have also had numbers and single letters removed. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

Fig. 5. Correlations between human judgments of paragraph similarity on the WENN dataset with estimates made using LSA (at 300 dimensions) using the WENN Wikipedia-based corpora containing 1000 and 10000 documents retrieved using Lucene with WENN-based query. Wikipedia documents have been truncated in four ways: first 100, 200, 300, and ALL words. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

Fig. 6. Correlations between human judgments of paragraph similarity on the Lee dataset with estimates made using LSA (at 300 dimensions) using Lee Wikipedia-based corpora containing 1000 and 10000 documents retrieved using Lucene with Lee-based query. Wikipedia documents have been truncated in four ways: first 100, 200, 300, and ALL words. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

Fig. 7. Correlations between human judgments of paragraph similarity on the WENN dataset with semantic model estimates made using Wikipedia Corpora with 1000 & 10000 documents and the WENN Corpus (NN-NSL). Error bars are the 95% confidence limits of the correlation. These results are also presented in Table J1 in the Supplementary Material file. Correlations exclude Same-Same paragraph comparisons.

Fig. 8. Correlations between human judgments of paragraph similarity on the Lee dataset with semantic model estimates made using Wikipedia Corpora with 1000 & 10000 documents and the Toronto Star (NN-NSL). Error bars are the 95% confidence limits of the correlation. These results are also presented in Table J2 in the Supplementary Material file. Correlations exclude Same-Same paragraph comparisons.

Fig. 9. Correlations between human and model estimates of paragraph similarity on the Lee dataset using the standard Wikipedia 1000 corpora (Wikipedia 1000) and Wikipedia 1000 corpora including the 50 Lee documents (Wikipedia 1050). The overlap model has also been included in this bar graph to allow the reader another point of comparison. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

Fig. 10. Correlations between human and model estimates of paragraph similarity on the Lee dataset using the standard Wikipedia 10000 corpora (Wikipedia 10000) and Wikipedia 10000 corpora including the 50 Lee documents (Wikipedia 10050). The overlap model has also been included in this bar graph to allow the reader another point of comparison. Error bars are the 95% confidence limits of the correlation. Correlations exclude Same-Same paragraph comparisons.

Fig. 11. Scatterplots of the two best similarity estimates calculated for both the WENN and Lee datasets compared to the average similarity estimates made by humans for each pair of paragraphs. On the WENN dataset, (A) LSA using the WENN corpus (NN-NSL), and (B) the Overlap model. On the Lee dataset, (C) Vectorspace using the Wikipedia 1050 (including Lee documents), and (D) the Overlap model. Note, on the Lee dataset, average human ratings have been normalized [0,1].

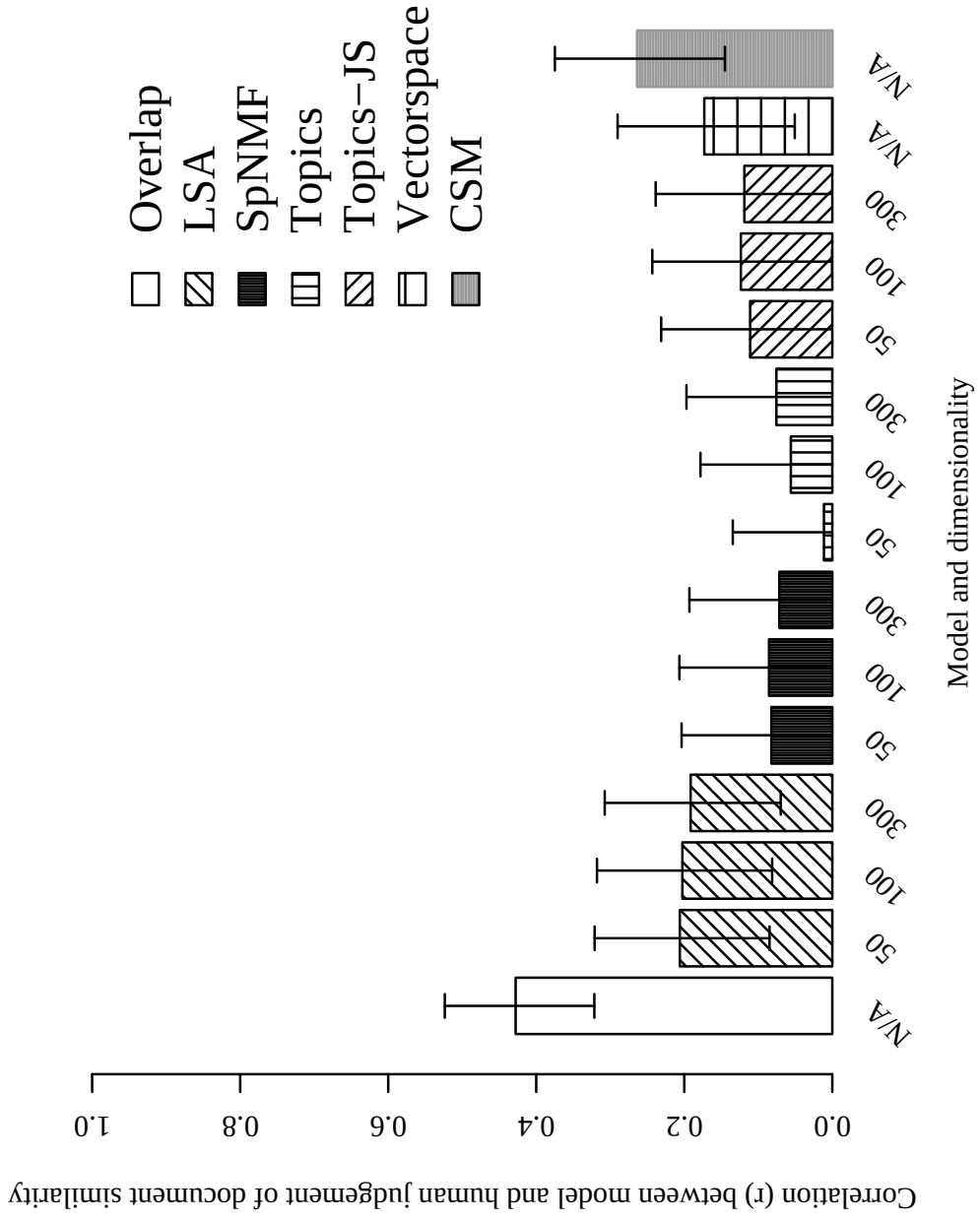

Overlap
LSA
SpNMF
Topics
Topics−JS
Vectorspace
CSM
Correlation (r) between model and human judgement of document similarity
0.0
0.2
0.4
0.6
0.8
1.0
N/A
50
100
300
50
100
300
50
100
300
50
100
300
N/A
N/A
Model and dimensionality

Fig. 1

Correlation (r) between model and human judgement of document similarity
0.0
0.2
0.4
0.6
0.8
1.0
N/A
50
100
300
50
100
300
50
100
300
50
100
300
N/A
N/A
Model and dimensionality
Overlap
LSA
SpNMF
Topics
Topics−JS
Vectorspace
CSM

Fig. 2

Correlation (r) between model and human judgement of document similarity
0.0
0.2
0.4
0.6
0.8
1.0
Overlap
LSA
SpNMF
Topics
Topics−JS
VecSpace
CSM
ALL
NN−NSL

Fig. 3

Correlation (r) between model and human judgement of document similarity
0.0
0.2
0.4
0.6
0.8
1.0
Overlap
LSA
SpNMF
Topics
Topics−JS
VecSpace
CSM
ALL
NN−NSL

Fig. 4

Correlation (r) between model and human judgement of document similarity
0.0
0.2
0.4
0.6
0.8
1.0
100
200
300
ALL
Words in Document
Wikipedia 1000 documents
Wikipedia 10000 documents

Fig. 5

Correlation (r) between model and human judgement of document similarity
0.0
0.2
0.4
0.6
0.8
1.0
100
200
300
ALL
Words in Document
Wikipedia 1000 documents
Wikipedia 10000 documents

Fig. 6

Correlation (r) between model and human judgement of document similarity
0.0
0.2
0.4
0.6
0.8
1.0
Overlap
LSA
SpNMF
Topics
Topics−JS
VecSpace
CSM
Semantic Models
Overlap
Wikipedia 1000 documents
Wikipedia 10000 documents
WENN

Fig. 7

Correlation (r) between model and human judgement of document similarity
0.0
0.2
0.4
0.6
0.8
1.0
Overlap
LSA
SpNMF
Topics
Topics−JS
VecSpace
CSM
Semantic Models
Overlap
Wikipedia 1000 documents
Wikipedia 10000 documents
Toronto Star

Fig. 8

Correlation (r) between model and human judgement of document similarity
0.0
0.2
0.4
0.6
0.8
1.0
Overlap
LSA
SpNMF
Topics
Topics−JS
VecSpace
CSM
Semantic Models
Overlap
Wikipedia 1000 documents
Wikipedia 1050 (with Lee) documents

Fig. 9

Correlation (r) between model and human judgement of document similarity
0.0
0.2
0.4
0.6
0.8
1.0
Overlap
LSA
SpNMF
Topics
Topics−JS
VecSpace
CSM
Semantic Models
Overlap
Wikipedia 10000 documents
Wikipedia 10050 (with Lee) documents

Fig. 10

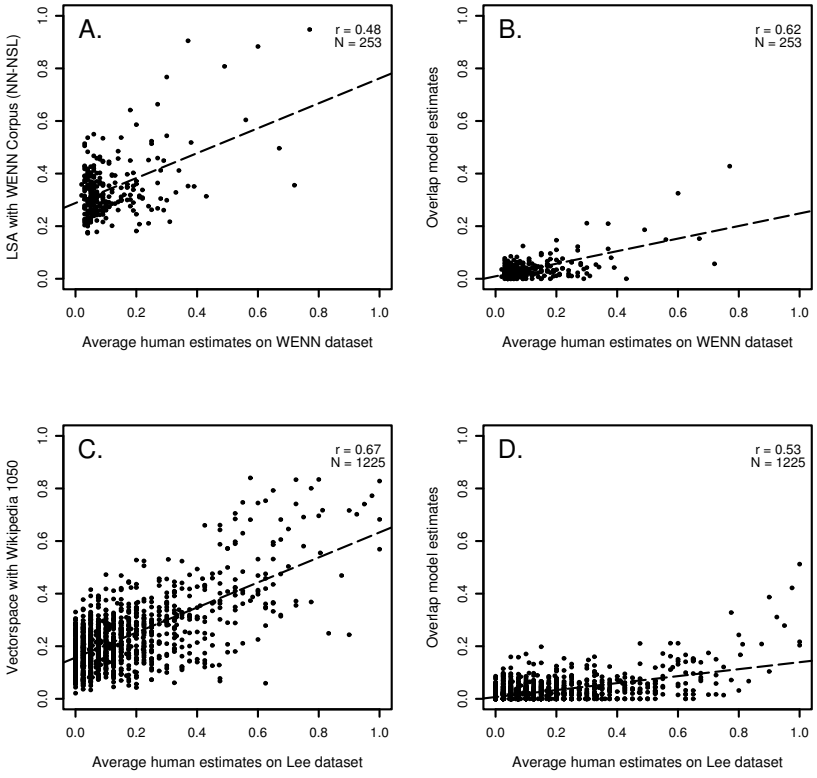

A.
r = 0.48
N = 253
LSA with WENN Corpus (NN-NSL)
Average human estimates on WENN dataset
B.
r = 0.62
N = 253
Overlap model estimates
Average human estimates on WENN dataset
C.
r = 0.67
N = 1225
Vectorspace with Wikipedia 1050
Average human estimates on Lee dataset
D.
r = 0.53
N = 1225
Overlap model estimates
Average human estimates on Lee dataset

Fig. 11

# Appendix J

## J. Paper 3 - Supplementary Material file

# Supplementary Material:

# Comparing Methods for Single Paragraph Similarity Analysis

## Appendix A

## Examples of similar and dissimilar paragraphs as rated by humans for the WENN dataset

*Similar paragraphs*

Paragraph 1: The woman accused of stalking Catherine Zeta-Jones will plead not guilty when she appears in court in two weeks time. Dawnette Knight is due to stand trial on November 10 on one charge of stalking and 24 charges of making criminal threats to the Chicago actress. Knight is currently being held in police custody, but her lawyers are appealing for bail, claiming she is only facing criminal proceedings because the case involves celebrities. Her lawyer Richard Herman says, “I’m afraid that we’ve gone much too far in something that’s just greatly overblown. She’s been in long enough. We all know that she’s no threat to anyone.” Welsh-born star Zeta-Jones, 35, who is married to actor Michael Douglas, said the “satanic threats” - which she received while she was filming Ocean’s Twelve in Amsterdam last year - left her on the verge of a nervous breakdown.

Paragraph 2: The American woman accused of stalking Catherine Zeta-Jones has written a letter of apology to the star. Dawnette Knight - who is being held on $1 million bail - has penned a note to the Oscar-winning actress, and her father-in-law Kirk Douglas, apologizing for any “distress” she may have caused. Knight, was arrested at her Beverly Hills home on June 3 and charged with one count of stalking and 24 counts of criminal threats. She admits she is obsessed with Zeta-Jones’ husband, Michael Douglas. Zeta-Jones’ lawyer Richard Herman released the letter which reads: “I was a confused young woman infatuated with Michael Douglas and have not rational explanations for my actions.” Her letter continues to explain she would “have never done anyone any harm and

would never harm anyone". She finishes: "It would be a wonderful good deed if you would all forgive me so that I can go back to college to finish my studies in child psychology." Knight, 32, is being held in custody and could receive a prison sentence of up to 19 years if convicted.

*Dissimilar paragraphs*

Paragraph 1: Movie superstar Tom Cruise has become the highest earning actor in Hollywood history after signing a deal that could earn him a staggering $360 million for his role in War Of The Worlds. Rather than agree a set fee for his part in the Steven Spielberg-directed epic, Cruise will earn 10 per cent of the film's box office takings plus a share of profits from DVDs, video games and toys. Experts predict the film - based on HG Wells' classic novel about a Martian attack - could make $1.8 billion at the cinema alone, of which Cruise's share would be an incredible $180 million. And, if he stars in the two planned sequels, Cruise's earnings will double at least. A Hollywood source says, "No expense will be spared. Spielberg wants to make it the film of the decade - the one that everyone talks about and rushes to see."

Paragraph 2: Superstar couple Victoria Beckham and David Beckham are desperate to add another child to their family in a bid to repair the damage done to their marriage by details of the soccer ace's alleged infidelity. The sexy pair recently canceled a planned promotional trip to America in favor of a two-week break in Morocco, and British newspaper The Sun reports they're using their intimate spell in exotic capital Marrakech to try for another baby. Friends of the couple claim they're ideally hoping to welcome a baby girl into the world to give sons Brooklyn, five, and 21-month-old Romeo a younger sister. The Beckhams' marriage became the subject of intense scrutiny earlier this year when David's ex-personal assistant Rebecca Loos revealed she'd enjoyed a steamy affair with

the Real Madrid player. A source says, “They have talked about having more children and would be thrilled if they had a little girl. There’s nothing either of them feel is more important than their kids - and David simply adores them. A new baby would be a great way for them to put their troubles behind them and start a new life together in Spain.”

## Appendix B

## Examples of similar and dissimilar paragraphs as rated by humans for the Lee dataset

*Similar paragraphs*

Paragraph 1: Nigerian President Olusegun Obasanjo said he will weep if a single mother sentenced to death by stoning for having a child out of wedlock is killed, but added he has faith the court system will overturn her sentence. Obasanjo's comments late Saturday appeared to confirm he would not intervene directly in the case, despite an international outcry.

Paragraph 2: An Islamic high court in northern Nigeria rejected an appeal today by a single mother sentenced to be stoned to death for having sex out of wedlock. Clutching her baby daughter, Amina Lawal burst into tears as the judge delivered the ruling. Lawal, 30, was first sentenced in March after giving birth to a daughter more than nine months after divorcing.

*Dissimilar paragraphs*

Paragraph 1: Very few women have been appointed to head independent schools, thwarting efforts to show women as good leaders, according to the Victorian Independent Education Union. Although they make up two-thirds of teaching staff, women hold only one-third of principal positions, the union's general secretary, Tony Keenan, said. He believed some women were reluctant to become principals because of the long hours and the nature of the work. But in other cases they were shut out of the top position because of perceptions about their ability to lead and provide discipline.

Paragraph 2: Beijing has abruptly withdrawn a new car registration system after

drivers demonstrated “an unhealthy fixation” with symbols of Western military and industrial strength - such as FBI and 007. Senior officials have been infuriated by a popular demonstration of interest in American institutions such as the FBI. Particularly galling was one man’s choice of TMD, which stands for Theatre Missile Defence, a US-designed missile system that is regularly vilified by Chinese propaganda channels.

## Appendix C

### Standard stop-list

The stop-list used in this research prior to the removal of single numbers and letters.

a about above across after afterwards again against all almost alone along already also although always am among amongst amoungst amount an and another any anyhow anyone anything anyway anywhere are around as at back be became because become becomes becoming been before beforehand behind being below beside besides between beyond bill both bottom but by call can cannot cant co computer con could couldnt cry de describe detail do done down due during each eg eight either eleven else elsewhere empty enough etc even ever every everyone everything everywhere except few fifteen fify fill find fire first five for former formerly forty found four from front full further get give go had has hasnt have he hence her here hereafter hereby herein hereupon hers herself him himself his how however hundred i ie if in inc indeed interest into is it its itself keep last latter latterly least less ltd made many may me meanwhile might mill mine more moreover most mostly move much must my myself name namely neither never nevertheless next nine no nobody none noone nor not nothing now nowhere of off often on once one only onto or other others otherwise our ours ourselves out over own part per perhaps please put rather re same see seem seemed seeming seems serious several she should show side since sincere six sixty so some somehow someone something sometime sometimes somewhere still such system take ten than that the their them themselves then thence there thereafter thereby therefore therein thereupon these they thick thin third this those though three through throughout thru thus to together too top toward towards twelve twenty two un under until up upon us very via was we well were what whatever when

whence whenever where whereafter whereas whereby wherein whereupon wherever whether which while whither who whoever whole whom whose why will with within without would yet you your yours yourself yourselves

## Appendix D

## Corpora Parameters

Parameters for each corpus used in this research are contain in Table D1.

---

Insert Table D1 about here

---

## Appendix E

## Study One results tables

*t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, with the WENN (see Table E1) and Lee (see Table E2) datasets generated by human raters in Study One.

---

Insert Table E1 about here

---

---

Insert Table E2 about here

---

## Appendix F

## Stop-list used by Pincombe 2004

Below is the stop-list used by Pincombe (2004) and Lee, Pincombe, and Welsh (2005). Also note that these researchers only included alphabetical characters (Pincombe, 2004, p. 14), therefore excluding both numbers and single letters from their corpora.

a about all also although am an and another any anybody anyhow anyone anything anywhere are as at b be become been being but by c can cannot could d did do does doing done e each eg either else et etc every ex f for from g h had has have having he hence her hers herself high him himself his how however i ie if in inc indeed is it its j k l ltd m many may me might more mr mrs ms must my myself n no nor not o of oh or otherwise ought our ours ourselves p per put q r re s self selves shall she should sl so some somehow such sup t than that the their theirs them themselves then there therefore these they this those though thus to u us v very via viz vs w was we were what whatever when whence whenever where whereafter whereas whereby wherein whereupon wherever whether which whichever while whither who whoever whole whom whose why will with within without would x y yes you your yours yourself yourselves z

## Appendix G

### Study Two result tables

Correlations (*r*) between similarity assessments of human raters and those made using LSA, Topic Model (Topics), Topic Model with Jensen-Shannon equation (Topics-JS), SpNMF at 50, 100, and 300 dimensions, and both the Vectorspace and CSM models (see Table G1). *t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, on the WENN (see Table G2) and Lee (see Table G3) datasets used in Study Two.

Insert Table G1 about here

Insert Table G2 about here

Insert Table G3 about here

## Appendix H

### IMDB-based Lucene query for Wikipedia

“naomi watts” OR “elevator” OR “fear” OR “liz* taylor” OR “eliz* taylor” OR “battling” OR “congestive” OR “heart” OR “failure” OR “whitney houston” OR “unhurt” OR “crash” OR “zeta jones” OR “stalker” OR “apologizes” OR “trial” OR “courtney love” OR “drug” OR “case” OR “gary busey” OR “jailed” OR “late” OR “jessica simpson” OR “nick lachey” OR “split” OR “george clooney” OR “contemplates” OR “legal” OR “action” OR “kirsten dunst” OR “broke” OR “jake gyllenhaal” OR “heart” OR “pals” OR “jennifer lopez” OR “mother” OR “pines” OR “affleck” OR “lindsay lohan” OR “health” OR “battle” OR “michael jackson” OR “lawyers” OR “accuser” OR “undergo” OR “katherine zeta jones” OR “stalker” OR “plead” OR “guilty” OR “richard gere” OR “horseriding” OR “accident” OR “britney spears” OR “attacks” OR “paparazzi” OR “mary kate olsen” OR “comeback” OR “vmas” OR “tom cruise” OR “penelope cruz” OR “reunite” OR “dinner” OR “macaulay culkin” OR “arrested” OR “dog” OR “trip” OR “victoria beckham” OR “david beckham” OR “plan” OR “baby” OR “number” OR “demi moore” OR “pregnant” OR “posh” OR “victoria beckham” OR “flees” OR “france” OR “intruder” OR “incident” OR “ben affleck” OR “jennifer garner” OR “linked” OR “cameron diaz” OR “justin timberlake” OR “fight” OR “paparazzi” OR “tom cruise” OR “million” OR “war worlds” OR “newlyweds” OR “jessica simpson” OR “nick lachey” OR “look”

## Appendix I

### Lee-based Lucene query for Wikipedia

(“australia*” and “democrats senator*” and “leader”) or (“amp” or “stock market”) or ((“mugabe” or “zimbabwe”) and “famine”) or (“alqaida” or “puk”) or ((“washington” or “us” or “usa” or “u s a” or “united states of america”) and (“georgian sovereignty” or “tbilisi”)) or ((“gay” or “homosexual” or “homo sexual”) and “discriminat*”) or (“saudi” and (“osama bin laden” or “bin laden” or “alqaida”)) or (“saddam hussein” or “abu nida”) ((“hunan” or “china”) and “flood”) or ((“warplanes” or “bombed”) and (“basra” or “iraq”)) or (“iraq” and “russia” and (“economic” or “cooperation”)) or (“saddam hussein” and (“weapons of mass destruction” or “wmd”)) or (“investigate” and (“taskforce” or “corruption”)) or ((“andrew bartlett” or “aden ridgeway” or “natasha stott despoja”) and “democrats”) or (“glass ceiling” or “woman s rights” or “equal opportunity”) or (“war” and “iraq” and (“bush” or “us” or “usa” or “united states of america” or “u s a”)) or ((“beijing” or “chinese”) and “government”) or ((“africa*” or “malawi” or “mozambique” or “zambia” or “angola” or “swaziland” or “lesotho” or “zimbabwe”) and (“starv*” or “faminine”)) or ((“malawi” or “africa*”) and (“hiv” or “aids”)) or ((“un” or “united nation*” or “u n”) and “environment”) or ((“fatah revolutionary council” or “frc”) and “terror*”) or (“work” and “dole”) or (“anthrax” and “biowarfare”) or ((“china” or “chinese”) and “missile”) or (“death” and “stoning”) or (“death” and “stoned”) or (“warheads”) or ((“fbi” or “federal bureau investigation”) and “terrorism”) or (“tamal tiger*”) or (“crim*” and “voyeur*”) or (“crim*” and “video*”) or ((“australia*” or “tampa”) and (“refugee” or “asylum”)) or (“australia” and “democrat*”) or (“whale” and “rescue”) or ((“prince william” or “price harry” or “princess di*”) and “ken wharfe”) or

((“osama bin laden” or “bin laden”) and (“jihad” or “holy war”)) or (“johannesburg earth summit”) or (“mugabe” and “zimbabwe”) or ((“men s rights” or “mens rights”) and “movement”) or (“global warming” or (“environment*” and “degradation”)) or (“bird*” and “tag*” and “research*”) or ((“un” or “u n” or “united nations”) and “sustainable growth”) or ((“russia*” or “ussr” or “u s s r”) and “chinese worker*”) or (“australia*” and “tampa”) or (“batasuna”) or ((“river” or “water”) and “flood*”) or ((“europe*” and “palestin*”) and (“isreal*” or “jew*”)) or (“job” and (“cuts” or “retrench*”)) or ((“asylum” or “refugee”) and “australia*”)

## Appendix J

## Study Three result tables

Human to model correlations when estimating similarity on the Wenn (see Table J1) and Lee (see Table J2) datasets, complex models using Wikipedia 1000 & 10000 document and domain-chosen corpora (without numbers or single letters – NN-NSL). Also, examples of the dimensions created by SpNMF on the 10000 document Wikipedia corpus generated for the Lee dataset (see Table J3).

Insert Table J1 about here

Insert Table J2 about here

Insert Table J3 about here

## Appendix K

## Study Four results tables

Comparisons of model performance when the 50 Lee dataset paragraphs are added to the Wikipedia 1000 (see Table K1) and 10000 (see Table K2) Wikipedia sub-corpora.

---

Insert Table K1 about here

---

---

Insert Table K2 about here

---

Table D1

Corpus parameters for the Toronto Star corpus, WENN corpus, and sub-corpora drawn from Wikipedia (1000 and 10000 documents) for both WENN and Lee datasets.

| | Docs (D) | Words (W) | W/D | Unique W (UW) | UW/D |
|---|---|---|---|---|---|
| WENN | 12,787 | 957,806 | 74.91 | 22,915 | 1.79 |
| WIKI-WENN | 1,000 | 2,420,436 | 2,420.44 | 48,508 | 48.51 |
| WIKI-WENN | 10,000 | 26,983,256 | 2,698.33 | 180,225 | 18.02 |
| Toronto Star | 55,021 | 14,070,668 | 255.73 | 96,975 | 1.76 |
| WIKI-Lee | 1,000 | 2,267,287 | 2,267.29 | 34,285 | 34.29 |
| WIKI-Lee | 10,000 | 18,135,603 | 1,813.56 | 164,271 | 16.42 |

Table E1

*t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, with the human ratings contained in the WENN dataset. None of the models' performance significantly improved when dimensionality was increased (alpha 0.05). Significant decreases in performance would be indicated by negative values equal to or greater than 1.96. So, in no case was increased dimensionality associated with significant decrements to model performance.

| Model | 50 | 100 |
|---|---|---|
| LSA-100 | -0.31 | |
| LSA-300 | -0.66 | -0.76 |
| Topics-100 | 1.13 | |
| Topics-300 | 1.17 | 0.48 |
| Topics-JS-100 | 1.05 | |
| Topics-JS-300 | 0.34 | -0.31 |
| SpNMF-100 | 0.39 | |
| SpNMF-300 | -0.9 | -1.34 |

Table E2

*t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, with human ratings contained in the Lee dataset. Significant decreases in performance would be indicated by negative values equal to or greater than 1.96. So, in no case was increased dimensionality associated with significant decrements to model performance.

| Model | 50 | 100 |
|---|---|---|
| LSA-100 | 2.29 | |
| LSA-300 | 2.30 | 2.28 |
| Topics-100 | -1.52 | |
| Topics-300 | 2.83 | 4.03 |
| Topics-JS-100 | 3.17 | |
| Topics-JS-300 | 2.30 | 1.60 |
| SpNMF-100 | 3.31 | |
| SpNMF-300 | 6.16 | 1.90 |

Table G1

Correlations (*r*) between similarity assessments of human raters and those made using LSA, Topic Model (Topics), Topic Model with Jensen-Shannon equation (Topics-JS), SpNMF at 50, 100, and 300 dimensions, and also the Overlap, Vectorspace and CSM models. The ALL columns display correlations based on corpora that contain both numbers and single letters (as used in Study One), conversely the NN-NSL columns are based on corpora with No Numbers and No single Letters (NN-NSL). Correlations exclude Same-Same document comparisons.

| | WENN | | LEE | |
|---|---|---|---|---|
| Model: | ALL | NN-NSL | ALL | NN-NSL |
| Overlap | 0.43 | 0.62 | 0.48 | 0.53 |
| LSA-50 | 0.21 | 0.38 | 0.04 | 0.10 |
| LSA-100 | 0.20 | 0.41 | 0.05 | 0.11 |
| LSA-300 | 0.19 | 0.48 | 0.06 | 0.12 |
| Topics-50 | 0.01 | 0.22 | 0.02 | 0.07 |
| Topics-100 | 0.06 | 0.22 | 0.01 | 0.07 |
| Topics-300 | 0.08 | 0.25 | 0.06 | 0.07 |
| Topics-JS-50 | 0.11 | 0.26 | 0.10 | 0.15 |
| Topics-JS-100 | 0.12 | 0.29 | 0.11 | 0.17 |
| Topics-JS-300 | 0.12 | 0.28 | 0.13 | 0.17 |
| SpNMF-50 | 0.08 | 0.35 | 0.09 | 0.15 |
| SpNMF-100 | 0.09 | 0.37 | 0.13 | 0.16 |
| SpNMF-300 | 0.07 | 0.43 | 0.14 | 0.17 |
| Vectorspace | 0.17 | 0.41 | 0.10 | 0.20 |
| CSM | 0.26 | 0.16 | 0.15 | 0.17 |

Table G2

*t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, with human ratings contained in the WENN dataset used in Study Two. All corpora have had single letters and numbers removed. Significant decreases in performance would be indicated by negative values equal to or greater than 1.96. In no case was increased dimensionality associated with significant decrements to model performance.

| Model | 50 | 100 |
|---|---|---|
| LSA-100 | 2.18 | |
| LSA-300 | 3.63 | 3.69 |
| Topics-100 | 0.16 | |
| Topics-300 | 0.58 | 1.12 |
| Topics-JS-100 | 1.22 | |
| Topics-JS-300 | 0.63 | -0.38 |
| SpNMF-100 | 1.87 | |
| SpNMF-300 | 2.2 | 1.7 |

Table G3

*t* values calculated using Williams' formula (T2) comparing within model correlations, where models have dimensionality or topics, with human ratings contained in the Lee dataset in Study Two. All corpora have had single letters and numbers removed. Significant decreases in performance would be indicated by negative values equal to or greater than 1.96. In no case was increased dimensionality associated with significant decrements to model performance.

| Model | 50 | 100 |
|---|---|---|
| LSA-100 | 3.16 | |
| LSA-300 | 2.98 | 2.53 |
| Topics-100 | 1.96 | |
| Topics-300 | 0.15 | -1.21 |
| Topics-JS-100 | 2.43 | |
| Topics-JS-300 | 1.33 | 0.16 |
| SpNMF-100 | 0.91 | |
| SpNMF-300 | 1.28 | 1.09 |

Table J1

Human to model correlations when estimating paragraph similarity on the WENN dataset, complex models using Wiki(pedia) 1000 & Wiki 10000 document corpora and the WENN Corpus (NN-NSL). Correlations exclude Same-Same paragraph comparisons.

| Model | Corpus | *r* | Lower CI (95%) | Upper CI (95%) |
|---|---|---|---|---|
| Word Overlap | N/A | 0.62 | 0.53 | 0.69 |
| LSA | Wiki 1000 | 0.34 | 0.23 | 0.45 |
| LSA | Wiki 10000 | 0.26 | 0.14 | 0.37 |
| LSA | WENN | 0.48 | 0.38 | 0.57 |
| SpNMF | Wiki 1000 | 0.36 | 0.25 | 0.47 |
| SpNMF | Wiki 10000 | 0.39 | 0.28 | 0.49 |
| SpNMF | WENN | 0.43 | 0.33 | 0.53 |
| Topics | Wiki 1000 | 0.14 | 0.01 | 0.26 |
| Topics | Wiki 10000 | 0.24 | 0.13 | 0.36 |
| Topics | WENN | 0.25 | 0.13 | 0.36 |
| Topics-JS | Wiki 1000 | 0.27 | 0.15 | 0.38 |
| Topics-JS | Wiki 10000 | 0.25 | 0.13 | 0.36 |
| Topics-JS | WENN | 0.28 | 0.17 | 0.39 |
| Vectorspace | Wiki 1000 | 0.34 | 0.23 | 0.44 |
| Vectorspace | Wiki 10000 | 0.35 | 0.24 | 0.45 |
| Vectorspace | WENN | 0.41 | 0.30 | 0.51 |
| CSM | Wiki 1000 | 0.11 | -0.01 | 0.23 |
| CSM | Wiki 10000 | 0.12 | 0.00 | 0.24 |
| CSM | WENN | 0.16 | 0.04 | 0.28 |

Table J2

Human to model correlations when estimating paragraph similarity on the Lee dataset, complex models using Wiki(pedia) 1000 & 10000 document corpora and the Toronto Star (NN-NSL) corpus. Correlations exclude Same-Same paragraph comparisons.

| Model | Corpus | *r* | Lower CI (95%) | Upper CI (95%) |
|---|---|---|---|---|
| Word Overlap | N/A | 0.53 | 0.49 | 0.57 |
| LSA | Wiki 1000 | 0.51 | 0.46 | 0.55 |
| LSA | Wiki 10000 | 0.39 | 0.34 | 0.43 |
| LSA | Toronto Star | 0.12 | 0.06 | 0.17 |
| SpNMF | Wiki 1000 | 0.53 | 0.49 | 0.57 |
| SpNMF | Wiki 10000 | 0.46 | 0.41 | 0.50 |
| SpNMF | Toronto Star | 0.17 | 0.11 | 0.22 |
| Topics | Wiki 1000 | 0.48 | 0.43 | 0.52 |
| Topics | Wiki 10000 | 0.43 | 0.38 | 0.47 |
| Topics | Toronto Star | 0.07 | 0.01 | 0.12 |
| Topics-JS | Wiki 1000 | 0.36 | 0.31 | 0.41 |
| Topics-JS | Wiki 10000 | 0.42 | 0.37 | 0.47 |
| Topics-JS | Toronto Star | 0.17 | 0.11 | 0.22 |
| Vectorspace | Wiki 1000 | 0.55 | 0.51 | 0.59 |
| Vectorspace | Wiki 10000 | 0.56 | 0.52 | 0.59 |
| Vectorspace | Toronto Star | 0.20 | 0.14 | 0.25 |
| CSM | Wiki 1000 | 0.08 | 0.02 | 0.13 |
| CSM | Wiki 10000 | 0.10 | 0.04 | 0.15 |
| CSM | Toronto Star | 0.17 | 0.12 | 0.22 |

Table J3

Examples of dimensions created by SpNMF on the 10000 document Wikipedia corpus generated for the Lee dataset where document length has been truncated at 100 words.

| Dimension 1 | | Dimension 2 | | Dimension 3 | |
|---|---|---|---|---|---|
| pollution | 0.78 | weapons | 0.42 | al | 0.81 |
| climate | 0.21 | biological | 0.34 | qaeda | 0.21 |
| change | 0.14 | bwc | 0.25 | bin | 0.17 |
| global | 0.14 | warfare | 0.21 | laden | 0.14 |
| environmental | 0.13 | germ | 0.15 | itihaad | 0.10 |
| warming | 0.13 | toxin | 0.13 | osama | 0.10 |
| overuse | 0.12 | pathogen | 0.13 | qaida | 0.09 |
| waste | 0.12 | stockpiling | 0.13 | group | 0.09 |
| greenhouse | 0.11 | incapacitate | 0.13 | abd | 0.09 |
| ipcc | 0.10 | organism | 0.13 | fadl | 0.08 |
| contamination | 0.09 | adversary | 0.13 | islamic | 0.08 |
| fossil | 0.09 | agreement | 0.13 | islam | 0.08 |
| resources | 0.08 | virus | 0.12 | militant | 0.08 |
| overpopulation | 0.08 | employment | 0.12 | terrorist | 0.07 |
| fuels | 0.08 | disease | 0.12 | abu | 0.07 |
| water | 0.08 | outlawed | 0.11 | islamist | 0.07 |
| wmo | 0.08 | causing | 0.11 | sunni | 0.06 |
| conservation | 0.08 | devastating | 0.11 | nashiri | 0.06 |
| deforestation | 0.07 | impact | 0.11 | ahmed | 0.05 |
| issues | 0.07 | kill | 0.11 | ali | 0.05 |

Table K1

Comparison of models performance with standard Wikipedia 1000 corpora (Wiki 1000) and Wikipedia 1000 corpora including the 50 Lee paragraphs (Wiki 1050), using correlations between human and model estimates of paragraph similarity on the Lee dataset. Correlations exclude Same-Same paragraph comparisons. Significance tests were performed using Williams' T2 formula.

| Model | Wiki 1000 | Wiki 1050 | diff | t | p |
|---|---|---|---|---|---|
| LSA | 0.51 | 0.6 | 0.09 | 7.65 | $< 0.05$ |
| Vectorspace | 0.55 | 0.67 | 0.12 | 9.08 | $< 0.05$ |
| Topics | 0.48 | 0.49 | 0.01 | 3.02 | $< 0.05$ |
| Topics-JS | 0.36 | 0.36 | -0.01 | -1 | n.s. |
| SpNMF | 0.53 | 0.56 | 0.03 | 2.66 | $< 0.05$ |
| CSM | 0.08 | 0.08 | 0 | 6.51 | $< 0.05$ |

Table K2

Comparison of models performance with standard Wikipedia 10000 corpora (Wiki 10000) and Wikipedia 10000 corpora including the 50 Lee paragraphs (Wiki 10050), using correlations between human and model estimates of paragraph similarity on the Lee dataset. Correlations exclude Same-Same paragraph comparisons. Significance tests were performed using Williams' T2 formula.

| Model | Wiki 10000 | Wiki 10050 | diff | t | p |
|---|---|---|---|---|---|
| LSA | 0.39 | 0.40 | 0.01 | 14.56 | $< 0.05$ |
| Vectorspace | 0.56 | 0.58 | 0.02 | 6.92 | $< 0.05$ |
| Topics | 0.43 | 0.49 | 0.06 | 7.18 | $< 0.05$ |
| Topics-JS | 0.42 | 0.40 | -0.02 | -2.41 | $< 0.05$ |
| SpNMF | 0.46 | 0.44 | -0.02 | -4.75 | $< 0.05$ |
| CSM | 0.10 | 0.10 | 0 | 6.57 | $< 0.05$ |

# Appendix K

## K. Paper 4 - Original Article

Stone, B. & Dennis, S. (2010). Semantic models and corpora choice when using semantic fields to predict eye movement on web pages.
*submitted to International Journal of Human-Computer Studies*