



ORBIT - Online Repository of Birkbeck Institutional Theses

Enabling Open Access to Birkbeck's Research Degree output

Grounding semantic cognition using computational modelling and network analysis

<https://eprints.bbk.ac.uk/id/eprint/40414/>

Version: Full Version

Citation: Ghose, Ajitesh (2019) Grounding semantic cognition using computational modelling and network analysis. [Thesis] (Unpublished)

© 2020 The Author(s)

All material available through ORBIT is protected by intellectual property law, including copyright law.

Any use made of the contents should comply with the relevant law.

[Deposit Guide](#)
Contact: [email](#)

Grounding Semantic Cognition Using
Computational Modelling and Network Analysis

Ajitesh Ghose

Centre for Cognition, Computation and Modelling
Department of Psychological Sciences

Thesis submitted in partial satisfaction of the requirements for the
degree of Doctor of Philosophy (Ph.D.) at Birkbeck,
University of London

December 2018

Abstract

The overarching objective of this thesis is to further the field of grounded semantics using a range of computational and empirical studies. Over the past thirty years, there have been many algorithmic advances in the modelling of semantic cognition. A commonality across these cognitive models is a reliance on hand-engineering “toy-models”. Despite incorporating newer techniques (e.g. *Long short-term memory*), the model inputs remain unchanged. We argue that the inputs to these traditional semantic models have little resemblance with *real* human experiences. In this dissertation, we ground our neural network models by training them with real-world visual scenes using naturalistic photographs. Our approach is an alternative to both hand-coded features and embodied raw sensorimotor signals.

We conceptually replicate the mutually reinforcing nature of hybrid (feature-based and grounded) representations using silhouettes of concrete concepts as model inputs. We next gradually develop a novel grounded cognitive semantic representation which we call *scene2vec*, starting with *object co-occurrences* and then adding *emotions* and *language-based tags*. Limitations of our scene-based representation are identified for more abstract concepts (e.g. *freedom*). We further present a large-scale human semantics study, which reveals small-world semantic network topologies are context-dependent and that *scenes* are the most dominant cognitive dimension. This finding leads us to conclude that *there is no meaning without context*. Lastly, *scene2vec* shows promising human-like context-sensitive stereotypes (e.g. gender role bias), and we explore how such stereotypes are reduced by targeted debiasing.

In conclusion, this thesis provides support for a novel computational viewpoint on investigating meaning - *scene-based grounded semantics*. Future research scaling scene-based semantic models to human-levels through *virtual grounding* has the potential to unearth new insights into the human mind and concurrently lead to advancements in *artificial general intelligence* by enabling robots, embodied or otherwise, to acquire and represent meaning directly from the environment.

Declaration

I declare that the work presented in this thesis is my own. Where it builds on other people's work or ideas this is clearly marked.

Acknowledgements

First and foremost, I would like to express my gratitude to Professor Richard P. Cooper, my supervisor and mentor. His expertise, availability and patience proved invaluable. He has provided guidance at key moments in my research while also allowing me to work independently. Without his support, this work would not exist.

Throughout my PhD, I have had the great fortune of working for my boss Steven Yule, Director of Research at Sky, who has provided me with immense flexibility and support and helped me grow into a leader in the commercial world of Research and Data Science. I would also like to thank Dr Andrew Haughton, Insights Director at Sky, my main stakeholder, who has been a great ally and coach. He appreciates and understands the challenges, nuances and inherent complexities of predicting human behaviours and intentions. Over the last five years at Sky, in my role as Head of Strategic Analytics, I have had the pleasure of leading a large department with 25+ researchers, data scientists and engineers. I am grateful to everyone in my team, but in particular to my friends and colleagues Kevin Connolly, Chris McLean, Hsueh Qu Li, Adam Ball, Jeny James-Charman and Cotton Ghose.

I am also indebted to the late Prof. Svend Østergaard from Aarhus University, who mentored me ten years ago and introduced me to computational and statistical modelling. I am also grateful to BRGS (Birkbeck Research Graduate School) for providing a wide range of doctoral training and development opportunities and for awarding me the first prize in their inaugural PhD poster competition.

I would also like to thank Prof. James McClelland and Prof. Lawrence Barsalou for providing guidance and feedback during two Embodied and Situated Language Processing conferences (ESLP 2014, 2018). Lastly, I would like to thank my beloved wife, Michelle Li. She has been tremendously supportive and understanding throughout this entire process and has made numerous sacrifices to help me get here.

Dedication

To my wife, Michelle Li.

Contents

1 Introduction: From Pixels to Meaning.....	1
1.1 Background: Grounding Meaning.....	1
1.2 Thesis Overview	3
2 Artificial Intelligence and Robotics: Origins of Grounded Cognition	8
2.1 Abstract	8
2.2 Introduction.....	9
2.3 Classical Theory of Cognition.....	10
2.4 Development of Grounded Cognition.....	13
2.4.1 Grounded Cognition: Roots in Philosophy and Linguistics.....	15
2.4.2 Grounded Cognition: Roots in Psychology	19
2.4.3 Grounded Cognition: Roots in AI and Robotics	20
2.4.4 Summary	27
2.5 Limitations of Grounded Artificial Intelligence.....	28
2.6 Towards a Computational Grounded Semantics	32
2.7 Summary.....	35
3 Feature-based and Grounded Semantic Representations.....	37
3.1 Abstract	37
3.2 Introduction.....	38
3.3 Disembodied Symbolic Models.....	40
3.4 Latent Semantic Analysis	45
3.5 Disembodied Sub-Symbolic Models.....	49
3.6 Grounded Developmental Robotics	54
3.7 Computational Study I: Noise Tolerance	59
3.7.1 Theoretical Background	59
3.7.2 Methodology.....	62
3.7.3 Results.....	66
3.8 Discussion	69
3.9 Summary.....	72
4 Extending Symbol Interdependency: Perceptual Scene Vectors	73
4.1 Abstract	73
4.2 Introduction.....	74
4.3 Evidence for Symbol Interdependency	78
4.4 Surface Semantic Analysis.....	81
4.5 Computational Experiment: Perceptual Scene Vectors	83
4.5.1 Theoretical Background	83
4.5.2 Experimental Hypotheses.....	90
4.5.3 General Methodology.....	91
4.5.4 Experiment 1: 20 concepts × 26 n-gram features.....	93
4.5.5 Experiment 2: 20 concepts × 26 random n-gram noun features	94
4.5.6 Experiment 3: 20 concepts × 26 random n-gram verb features.....	95
4.5.7 Experiment 4: 20 concepts × 300 / 150 / 50 LSA dimensions	97

4.5.8	Experiment 5: 20 concepts × Perceptual Scene Vectors (PSVs).....	98
4.5.9	Experiment 6: 20 new concepts × 300 LSA dimensions	106
4.5.10	Experiment 7: 20 new concepts × Perceptual Scene Vectors (PSVs)	107
4.5.11	Quantifying Semantic Coherence	109
4.6	Discussion	113
5	Grounding Concrete versus Abstract Semantics	120
5.1	Abstract	120
5.2	Introduction.....	121
5.2.1	Comparing Concrete and Abstract Words.....	122
5.2.2	Modelling Abstract Semantics	124
5.2.3	Dimensions of Semantic Representations	127
5.2.4	Concreteness Revisited	135
5.2.5	Role of Emotions in Grounding Abstract Concepts.....	136
5.3	Computational Study I: Evaluating PSVs using BrainBench.....	139
5.3.1	Objective.....	139
5.3.2	Methodology.....	140
5.3.3	Results.....	141
5.4	Computational Study II: Grounding Concrete to Abstract Concepts.....	142
5.4.1	Objective.....	142
5.4.2	Methodology.....	142
5.4.3	Results.....	143
5.5	Computational Study III: Extending PSVs with Emotions.....	145
5.5.1	Objective.....	145
5.5.2	Methodology.....	147
5.5.3	Results.....	150
5.6	Discussion	155
6	Network Topology of Semantics: Grounding and Relativity of Meaning.....	164
6.1	Abstract	164
6.2	Introduction.....	165
6.3	Multidimensional Semantic Space	167
6.4	Human Experiment: Mapping our Meaning Space	174
6.4.1	Visualising Semantic Space	174
6.4.2	Relative Importance of Dimensions	176
6.4.3	Brain-based Componential Semantics	178
6.4.4	Relativity of Meaning	183
6.4.5	Objectives	185
6.4.6	Methodology.....	191
6.4.7	Results.....	198
6.5	Discussion	229
6.5.1	Contributions.....	229
6.5.2	Limitations	237
6.5.3	General Implications	239
6.5.4	Implications for AI.....	243
6.5.5	Summary	244
7	Gender Bias in Grounded Semantics: Network Regularisation and Debiasing	246
7.1	Abstract	246
7.2	Introduction.....	247
7.2.1	Historical and Empirical Foundations of Bias	249

7.2.2	Implicit Gender Biases	250
7.2.3	Biases in Machine Learning.....	251
7.2.4	Gender Biases in Machine Learning.....	254
7.2.5	Awareness of Debiasing Machine Learning	254
7.3	Investigating Gender Bias in Grounded Semantics.....	255
7.3.1	Objective.....	255
7.3.2	Methods.....	256
7.3.3	Results.....	265
7.4	Discussion.....	282
7.4.1	Contributions.....	283
7.4.2	Importance of Grounded Data.....	285
7.4.3	Gender Bias.....	289
7.4.4	Limitations	289
7.4.5	Debiasing Semantic Networks.....	291
7.4.6	Summary	297
8	General Discussion	299
8.1	Abstract	299
8.2	Introduction.....	300
8.3	Main Contributions	302
8.4	Responding to Critiques of Grounded Cognition	305
8.4.1	Grounded Semantics and Dreyfus' Critiques.....	305
8.4.2	Grounded Semantics and Johnson-Laird et al.'s Critique	309
8.5	Future Applications.....	310
8.5.1	Implications for Artificial Intelligence and Robotics	310
8.5.2	Implications for Advertising and FinTech	315
8.6	Future Research	319
8.6.1	Empirical Ground Truths for Semantic Modelling.....	319
8.6.2	Large-scale Cognitive Semantic Modelling.....	320
8.6.3	Development of Semantic Network Topologies	322
8.7	Conclusion	323
Appendix A	Network Analysis	325
A.1	Overview	325
A.2	Graph Theory.....	325
A.3	Network Analysis.....	326
A.4	Main Network Metrics	327
Appendix B	Limits of Simple Plots	329
B.1	Overview.....	329
B.2	Challenges of Visualising Multiple Dimensions.....	329
Appendix C	Limits of PCA and MDS.....	331
C.1	Overview.....	331
C.2	Dimensionality Reduction.....	331
Appendix D	Enlarged Images of Results.....	333
References		349

List of Figures

Figure 1.1 (A) The traditional view of cognition playing a mediating role between perception and action. (B) Brooks' model reveals cognition as an emergent property of perception and action.....	2
Figure 2.1: Example of a Physical Symbol System.....	12
Figure 2.2: Comparison of traditional and embodied symbol states	13
Figure 3.1: Arbor porphyriana (tree of Porphyry)	44
Figure 3.2: Example format of the feature-based representations.....	62
Figure 3.3: Silhouettes and their corresponding silhouette profiles based on the 2D to 1D transformation	64
Figure 3.4: Performance of the neural networks trained on feature-based, grounded and hybrid representations as a function of noise	67
Figure 3.5: Classification accuracies of the grounded neural network	68
Figure 3.6: A Multidimensional Scaling (MDS) representation of the grounded semantic space (left), and a hierarchical cluster plot of the same grounded semantic space (right) ..	69
Figure 4.1: Charles Sanders Peirce's notion of Sign.....	75
Figure 4.2: MDS plots from Louwerse (2011) of the 16 verbal descriptors \times 26 features used in Rogers and McClelland (2004).....	83
Figure 4.3: The differentiation of conceptual representations based on 26 hand-coded n-gram features (a + d), 26 random noun n-gram features (b + e) and 26 random verb n-gram features (c + f).....	96
Figure 4.4: The differentiation of conceptual representations based on 300 LSA dimensions (a + d), 200 LSA dimensions (b + e) and 100 LSA dimensions (c + f).....	98
Figure 4.5: Zhao et al.'s PSPNet framework optimised for extracting global-scene-level priors.....	101
Figure 4.6: PSPNet's 3 level of abstraction: (i) objects (level 1), (ii) object parts (level 2), and (iii) parts of object parts (level 3)	102
Figure 4.7: Example from Zhou et al. demonstrating object-level scene segmentation....	103
Figure 4.8: Schematized example of steps 3 and 4, respectively, extracting the probabilities and binarising.....	103
Figure 4.9: Conceptual representations of 20 verbal descriptors with discernible taxonomic and associative hierarchies based on the hidden layer associations of the neural network trained using PSVs.....	106

Figure 4.10: (a + b) Conceptual representations of 20 novel animate and inanimate verbal descriptors based on 300 LSA-dimensions. (c + d) The relationships of the same 20 concepts based on the hidden layer associations of the simple feedforward neural network trained using PSVs.....	108
Figure 5.1: The recurrent neural network from Hoffman et al. (2018).....	125
Figure 5.2: (A) A histogram of the concreteness ratings of 40,000 word lemmas originating from Brysbaert et al. (2014) (B) A boxplot of the standard deviations of the same concreteness ratings	136
Figure 5.3: A bar chart of the BrainBench results across a range of distributed “off-the-shelf” representations and our perceptual scene vector (PSV).....	141
Figure 5.4: A correlation plot of the PSV’s hidden layer representations. Concepts are grouped into concrete (blue), intermediate (green), and abstract (red) groupings.....	144
Figure 5.5: Example output of the emotion detection and the resultant JSON file.....	147
Figure 5.6: Schematic overview of PSVs, scene2vec and investigating the correspondence with LSA by correlating their respective representational dissimilarity matrices (RDMs)	149
Figure 5.7: Outline of the 12 different comparisons between grounded and language-based representations, using our adapted <i>representational similarity analysis</i> (RSA)	150
Figure 5.8: A correlation plot of the scene2vec’s hidden layer representations	152
Figure 5.9: Hierarchical cluster plot of the hidden layer neurons representing the semantic associations of PSVs (A) and scene2vec (B) representations	153
Figure 5.10: Concept- and category-level correlations of PSV and scene2vec representations in relation to LSA 300 distributed representations	154
Figure 6.1: Example of Troche et al.’s (2014) dichotomous measure of concreteness, with the x-axis depicting concreteness ratings.....	169
Figure 6.2: The three-dimensional semantic space generated by Troche et al. (2017).....	171
Figure 6.3: Visualising the semantic space generated using PCA of the 400 noun concepts in Lynott and Connell’s (2013) modality exclusivity norm study.....	178
Figure 6.4: Binder et al.’s (2016) cosine-similarity comparison of 434 nouns using brain-based representations (left) with LSA representation (right).	181
Figure 6.5: A screenshot of how the conceptual feature rating exercise looks like on a computer screen	194
Figure 6.6: A screenshot of how the MaxDiff exercise looks like on a computer screen. ..	195
Figure 6.7: Correlations between the 16 cognitive dimensions and the concreteness ratings merged from Brysbaert et al. (2014)	200
Figure 6.8: (A) Factor Analysis components matrix with conditional formatting (blue for larger numbers and red for smaller numbers). (B) A scree plot of the first nine principal components extracted.....	201

Figure 6.9: Comparing t-SNE's semantic embedding space (A) with MDS' space (B) for all 544 concepts	203
Figure 6.10: A matrix of network topologies as <i>t</i> -SNE's <i>perplexity</i> (columns) increases and the <i>correlation threshold</i> (rows) decreases, labelled A to P	205
Figure 6.11: Three views of the semantic topology of 544 concepts, based on nodes colour-coded to represent (A) the concreteness spectrum based on ratings integrated from Brysbaert et al. (2014), (B) the most dominant cognitive dimension selected in the MaxDiff discrete choice modelling task, and (C) the k-core clusters.....	207
Figure 6.12: Plotting the relationship between a log-rescaled t-SNE perplexity and (A) network components, (B) diameter, (C) clustering coefficient and (D) the small world index (SWI)	210
Figure 6.13: Overall semantic topology, including concept labels and the nodes of the network colour-coded according to the concreteness spectrum, ranging from red (abstract) to green (intermediate) and blue (concrete)	212
Figure 6.14: A random topology based on shuffling the semantic dimensions across 16 dimensions	217
Figure 6.15: (A) An adapted <i>Lorenz-style</i> plot of the cumulative distribution of network dimensions' maximal activations across the 19 k-core network clusters. (B) An overview of the 16 cognitive dimensions plotted based on the Gini coefficient and the <i>absolute semantic dominance</i> - a measure of the number of nodes being maximally activated.....	219
Figure 6.16: (A) Sankey plot of the incidence matrix with <i>place</i> -to-cluster relations highlighted. (B) Sankey plot of the incidence matrix with <i>ingestion</i> -to-cluster relations highlighted.....	221
Figure 6.17: (A) Semantic network in context-free general condition. (B) Semantic network in the home move context. (C) Semantic network in the cooking context. (D) Comparison of network degree distributions using density functions.....	223
Figure 6.18: (A) Semantic network in context-free general condition. (B) Semantic network in the house on fire context. (C) Semantic network in the water buoyancy context. (D) Comparison of network degree distributions using density functions.....	224
Figure 6.19: (A) Semantic network in context-free <i>general</i> condition. (B) Semantic network in the <i>car boot sale</i> context. (C) Semantic network in the <i>gifting</i> context. (D) Comparison of network degree distributions using density functions.....	225
Figure 7.1: Depiction of four concepts, along with an example image and automatically generated tags.....	259
Figure 7.2: Visualisation of the semantic networks grounded in Google Images.....	265
Figure 7.3: Visualisation of the semantic networks grounded in Getty Images.....	268
Figure 7.4: Three different network centrality measures (Betweenness, Closeness and Strength) across the Getty Images and Google Images networks.	269

Figure 7.5: Visualisation of the regularised semantic networks by applying a <i>graphical LASSO</i> on the network associations grounded in <i>Google Images</i> (A) and <i>Getty Images</i> (B)	272
Figure 7.6: A depiction of a small and quasi-representative set of images for the search term <i>nurse</i> in both <i>Google Images</i> (A) and <i>Getty Images</i> (B).	273
Figure 7.7: Quantifying the gender bias in the semantic networks grounded in <i>Google Images</i> (A) and <i>Getty Images</i> (B)	275
Figure 7.8: The regularised graphical LASSO networks are shown for the conditions consisting of debiasing (A) only <i>occupation</i> , (B) only <i>business</i> and (C) <i>occupation</i> and <i>business</i>	279
Figure 7.9: A summary plot of average bias levels for the original <i>scene2vec Google Images</i> network along with the three debiasing experimental conditions of (i) <i>occupation-only</i> , (ii) <i>business-only</i> and (iii) <i>occupation and business</i>	281
Figure 8.1: Depiction of two virtual scenes in <i>Real Sim City</i> modelled on real-world locations	311

List of Tables

Table 3.1: Model parameters and sum square error for ANN training of feature-based, grounded and hybrid models	65
Table 4.1: 20 verbal descriptors \times 26 features used in Rogers and McClelland (2004); Louwerse (2011) used the first 16 verbal descriptors (excluding mammals).....	92
Table 4.2: Summary of our predictions for experiments 1 through 7. Weak predictions indicate very poor associative and taxonomic relationships, whereas strong predictions more meaningful relationships.....	109
Table 4.3: Summary of Adjusted Rand Index (ARI) across all 7 experiments, quantifying the similarity in semantic clustering between the various language and Perceptual Scene Vector (PSV) experiments and the ground truth clustering from Rogers and McClell	112
Table 6.1: Troche et al.'s (2017) <i>conceptual feature ratings</i> (CFR) instrument, the first column contains the 14 cognitive dimensions and the second column the verbal descriptions for participants	170
Table 6.2: A summary of our central hypotheses	190
Table 6.3: Our conceptual feature ratings (CFR) instrument, based on Troche et al.' (2017) question format of "I relate this word to/with..."	193
Table 6.4: Stimuli used in the context-specific experiments (phase 3)	196
Table 6.5: Interclass correlations (ICC) of the 16 cognitive dimensions for both the <i>conceptual feature ratings</i> (CFR) and the <i>Maximum-Difference</i> (MaxDiff) scores.....	199
Table 6.6: The top-6 concepts (out of 544) for each of the 16 dimensions. The numerical score is the aggregated conceptual feature rating (CFR) measured on a 7-point Likert scale... ..	202
Table 6.7: Network metrics for SemNet and bootstrapped RandNet, including 95% confidence intervals for the null model (LCI: lower confidence interval, UCI: Upper confidence interval).	216
Table 6.8: The incidence matrix for calculating Gini coefficients.....	218
Table 6.9: Key network metrics across three context comparison sets, each containing a general context and two context-specific networks.....	226
Table 6.10: Summary of the <i>differential activation quotient</i> (DAQ) between the six grounded scenarios and the general context-free conceptual feature ratings (CFR)	227
Table 7.1: The 60 concepts used in the gender bias experiments. The concepts are grouped into ten categories.	257

Chapter 1

Introduction: From Pixels to Meaning

1.1 Background: Grounding Meaning

Thirty-five years ago, Rodney Brooks, a robotics pioneer from Stanford ran a popular lecture series titled “*From Pixels to Predicates*”, which provided a grand new vision of artificial intelligence and human intelligence. At the time, Brooks was only starting out as a young post-doc. Sixteen years later, in an influential review of the ground-breaking work, Brooks (1999) recalls the two simple overheads presented during this lecture series (see *figure 1.1*). The first (*figure 1.1A*) was a diagram depicting *traditional cognition* with cognition being a qualitatively distinct module from *perception* and *action*, while in the second diagram (*figure 1.1B*), cognition emerged dynamically from the coupling of perception and action. Therefore, cognition was not merely a module transforming sensorimotor information. This original perspective became popularised as *Brooks-style AI* and led to decades of research on how intelligence in *planning* and

decision making could be acquired dynamically through sensorimotor interactions in the real world.

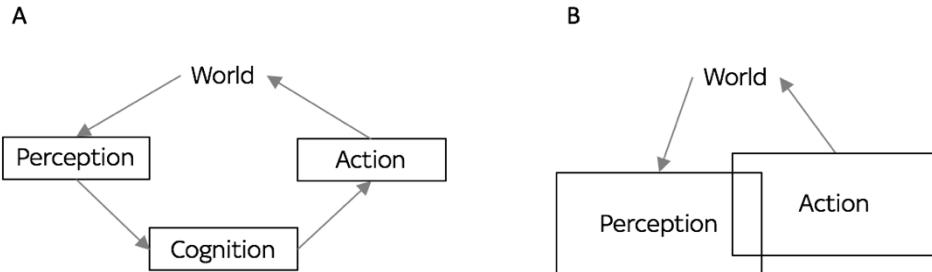


Figure 1.1 (A) The traditional view of cognition playing a mediating role between perception and action. (B) Brooks' model reveals cognition as an emergent property of perception and action. Source: Digital adaptations from Brooks (1999).

Despite the successes and (eventual) failures of pure Brooks-style AI, Brooks' original vision of grounding *predicates in pixels* (the visual world) was never realised in the field of *cognitive modelling* research. Traditional computer vision in the 1990s and even early 2000s were too cumbersome, inefficient and fragile to help accomplish Brooks' "pixels to predicates" research vision. Brooks (1999, p. ix) summarised this dilemma in the following passage:

"Computer vision systems ought to be able to operate in the ordinary sorts of environments that people operated in, cluttered offices with things stuck on walls and disorderly piles of papers that partially obscured objects...a computer vision system should be able to operate outdoors and pick out trees, hills, pathways, curbs, houses, cars, trucks, and everything else that a three-year-old child could name. There were no vision systems around doing anything even remotely as sophisticated."

However, over the last decade, with significant advances in *deep learning* this is no longer the case (see LeCun, Bengio, & Hinton, 2015, for a brief review). Nonetheless, the most sophisticated and recent models of *cognitive semantics* (e.g. Rogers & McClelland, 2004; Hoffman et al., 2018) still rely on hand-coded, modeller-defined and idealised "toy-environments" which have little in common with human

phenomenological experiences of our environment. In this dissertation, we are inspired by Rodney Brooks' influences and the increasing importance of grounded theories of cognition in psychology, despite a lack of mechanistic models of grounded semantics. We propose to fill this gap in the extant literature through our present research programme on grounding meaning in the real-world complexities of naturalistic visual scenes.

1.2 Thesis Overview

Here, we briefly provide an overview of the seven core chapters constituting the present dissertation, primarily based on *artificial neural network* modelling, *network analysis* and one large-scale *human study* on *context-dependent* and *-independent* semantics with over 500 concepts and more than 2,000 participants in an online survey.

In chapter 2, we review the extant literature on grounded cognition and outline its increasingly influential role in emphasising the importance of real-world sensorimotor associations constituting the building blocks of intelligence. It rejects the notion of cognition being the result of computational manipulation of *amodal symbols*. We discuss widely cited contemporary reviews on grounded cognition by psychologists like Lawrence Barsalou, which frequently overlook the historical influences of *artificial intelligence* (AI) and *cognitive robotics* while only sketching out the importance of these fields for the future of grounded cognition research. In our review, in addition to highlighting well-known philosophical, linguistic and empirical influences, we also incorporate overlooked Asian philosophical precursors and focus on the origins rooted in AI and robotics research from the 1970s and 1980s as an alternative paradigm to *GOFAI* (*Good Old-Fashioned AI*). Thus, essential but seemingly neglected influences from AI and robotics on modern-day grounded cognition are resurfaced. We conclude with an overview of the importance of semantics as a testbed

for grounded cognition and discuss future implications for scaling computational models of human and artificial systems situated in the real world.

In chapter 3, we conceptually replicate the findings from Goldstone and Rogosky's (2002) study based on the ABSURDIST graph-matching algorithm. Fodor (1998) claimed that relations between concepts in a semantic system are insufficient for mapping correspondences between concepts. The ABSURDIST model, using relative and absolute distance measures between concepts, shows that relations between concepts are critical for semantic representations. Our investigation differs from Goldstone and Rogosky's in one critical manner. In-line with the present thesis of real-world grounding, we opt not to use the technique of generating concepts using an arbitrary set of *absolute* (extrinsic) and *relative* (intrinsic) coordinates. Instead, we use the well-known *feature-based* approach from Rogers and McClelland (2004) for encoding *distributed* representations while using our novel 2D silhouette-based technique for encoding *extrinsic* or grounded representations of concept images. The same neural network architecture is implemented across all three conditions (*feature-based*, *grounded* and *hybrid*), where hybrid representations consist of half of the most informative features from both the feature-based and grounded input datasets. Our results support the findings of Goldstone and Rogosky (2002), by revealing that hybrid representations show a markedly slower rate of decline in classification accuracy for our concepts as a function of increasing levels of noise perturbations applied to the hidden layer representations. Grounded representations perform the poorest (highest rate of noise-intolerance), while feature-based inputs are moderately tolerant to increasing levels of noise. Therefore, hybrid representations (*grounded/extrinsic* and *distributed/intrinsic*) appear to be mutually reinforcing and superior to either grounded or feature-based representations in isolation.

In chapter 4, we review Louwerse's (2011) experiments supporting the *symbol interdependency hypothesis*, which posits that meaning extraction attributed to embodied representations or algorithms should instead be assigned to language. In a range of computational experiments, we find evidence for language surface structures encoding meaning best when sufficiently constrained by modeller-determined feature sets, with performance deteriorating for randomly selected language surface structures. Furthermore, *Latent Semantic Analysis'* meaning encoding improves as weaker dimensions are removed. These findings collectively indicate that although language is important, increasing the relevance of linguistic, statistical regularities is also critical. Our novel approach, *Perceptual Scene Vectors* (PSVs), uses object co-occurrences from images to automatically extract strong associative and taxonomic relationships more successfully, measured both qualitatively and quantitatively, with an original application of a cluster-correspondence metric. PSVs encode meaning without modellers hand-coding relevant features, which provides an ecologically valid approach to extending symbol interdependency beyond language and partially solving the *relevance problem* in semantics by grounding meaning extraction in real-world visual scenes.

In chapter 5, we outline how empirical and computational studies on semantics have been limited to concrete concepts, despite the importance and prevalence of abstract words in the human lexicon. Recently, there has been an increased focus on describing the content of abstract concepts through *introspective*, *emotional*, *metaphorical* and *situational* descriptions (Borghi et al., 2017). The literature converges on *emotions* being essential for abstract words, hypothesised as *embodied abstract semantics* (Kousta et al., 2011). Here, we replicate the *concreteness continuum* by re-analysing data from a large-scale normative study (Brysbaert, Warriner, & Kuperman, 2014) and externally validate PSVs using a neuroimaging benchmark. We then compare situationally grounded concepts with traditional language-based representations and

find PSVs can successfully represent *concrete* but neither *intermediate* nor *abstract* concepts. Lastly, we develop our new *scene2vec* representation by extending PSVs with emotion labels extracted from photographs, which yield noticeably enriched semantic representations across the concreteness spectrum, despite a lower performance for more abstract concepts. Our original contribution of modelling semantics using emotions only partially supports the *embodied abstract semantics* hypothesis and indicates that there is more to representing abstract meaning than emotions alone.

In chapter 6, our large-scale human study, we collect semantic dimension and importance ratings from 2,062 participants on 544 English words spanning the concreteness spectrum. Critically, we also have six context-specific conditions (e.g. *imagine you are cooking*) followed by ratings of the same semantic dimensions. We generate a network topology using non-linear dimensionality reduction (*t-SNE*). Our novel application of graph-theoretical techniques to cognitive semantic networks reveals that (i) semantic networks have a *small-world structure*, (ii) context-free semantic networks are organised lexically on a concreteness gradient, (iii) changes in context can dynamically modulate the lexical network topology, and (iv) *scenes* are the most vital and influential dimension shaping our conceptual networks. Collectively, these findings support a grounded perspective on meaning. *There is no meaning without context.*

In chapter 7, we build on the results of object-co-occurrences capturing semantics (chapter 4), limitations discovered in exclusively scene-based meaning (chapter 5) and the dynamic topology of meaning captured using network analysis grounded in cognitive dimensions (chapter 6). Here, we extend *scene2vec* with an additional “off-the-shelf” algorithm linking visual scenes to *word tags*, based on our simulation results from chapter 3 which reveals the mutually reinforcing nature of amodal and modal representations. This modified *scene2vec* representation is used to explore the critical issue of *gender bias* in two popular web-based image repositories - *Google Images* and *Getty Images*. Using network analysis and

our novel application of *graphical LASSO regularisation* to computational semantics, we show, as predicted, that *scene2vec* trained on *Google Images*, but not on *Getty Images*, encodes well-established *gender-occupation* stereotypes from the psychological literature (e.g. *man-doctor* or *woman-nurse*). The presence of context-specific human-like gender biases in *scene2vec* provides a new and scalable method for mechanistically investigating bias. Lastly, we develop a simple debiasing technique, called *semantic feature neutralisation* (SFN), which can selectively target and remove undesirable biases in our small-scale semantic model, while leaving desirable gender associations intact (e.g. *man-trousers* or *women-skirt*).

Finally, in chapter 8 we conclude the dissertation with a summary of *why* grounded representations are essential for cognitive science and artificial intelligence. We outline three future research opportunities, which are: (i) creating a *ground truth* for benchmarking cognitive semantic models, (ii) large-scale semantic modelling, and (iii) developmental investigations of semantic network topologies. Lastly, we propose that a move towards human-level AI will benefit from realistic *virtual grounding*, enabling rapid iterative progress while also providing a novel, ecologically-valid foundation for cognitive modelling of semantics.

Chapter 2

Artificial Intelligence and Robotics: Origins of Grounded Cognition

2.1 Abstract

Grounded cognition is an increasingly influential theory of intelligence emphasising the importance of real-world sensorimotor associations constituting the building blocks of thought. It rejects the notion of cognition being the result of computational manipulation of *amodal symbols*. However, literature reviews of *grounded cognition* (e.g. Barsalou, 2008, 2010) typically focus on Western philosophical origins and an overview of the linguistic roots underpinning contemporary ideas of grounding. Additionally, these accounts frequently conclude with a summary of a range of promising current and future research in domains of *artificial intelligence* (AI) and *cognitive robotics*. In our brief review, we highlight some of these well-known philosophical, linguistic and empirical influences but also incorporate overlooked Asian philosophical precursors. Our original contribution consists of discussing the origins of grounded

cognition rooted in AI and robotics research from the 1970s and 1980s as an alternative paradigm to *GOFAl*. In addition to covering *behaviour-based robotics* (Brooks, 1986, 1989, 1990, 1991a, 1991b, 1991c) we also review less well-known grounded AI models (Munson, 1971) and *conceptual graphs* (Sowa, 1976, 1979). We conclude with an overview of the importance of semantics as a testbed for grounded cognition and future implications for scaling computational models of human and artificial systems situated in the real world.

2.2 Introduction

Cognitive science is the field dedicated to descriptively, inferentially and mechanistically understanding predominantly human mental processes. *Artificial intelligence*, on the other hand, is founded on the principles of developing useful real-world intelligent systems and studying human intelligence (Winston, 1984). In traditional frameworks of both human and artificial intelligence, theorists have viewed *cognition*, the mental processes constituting thought, as disembodied symbol manipulation. In this framework, abstract symbols are manipulated by rules and operators which collectively give rise to the full range of complex human meaning and behaviours. The German mathematician and philosopher Gottfried Wilhelm Leibniz (1646 - 1716) is one of the forefathers of this approach, based on the ambition of creating a logical calculus of all human concepts (Sun, 2008). More recently, Chomsky (1957), outlined a quintessentially symbolic model of language processing known as *transformational-generative grammar* focusing on *syntax, morphology* and *phonology* with detailed information processing steps leading to the emergence of linguistic patterns and subsequently, conceptual processing. Chomsky's work marked a significant departure from more traditional linguistic perspectives that overlooked explanations relating to the mind or the brain.

Grounded cognition (GC) is usually negatively defined as the opposite view towards this classical paradigm of intelligence. In GC, both the *environment* and the *body* are critical constituents to shaping cognition (Barsalou, 2010). Some theorists of grounded cognition (e.g. Froese, 2007; Chemero, 2011; Vallet, 2015; Varela, Thompson, & Rosch, 1991) have suggested GC leading a paradigmatic shift away from more traditional symbolic views to the ones incorporating the influence of our body and situational factors. Somewhat surprisingly, some even claim GC is the dominant new cognitive science paradigm (Stewart et al., 2010).

In this review, we start with an overview of so-called *classical cognition* and attempt to delineate its fuzzy boundary with grounded cognition. Our review of GC's extant literature is chronologically structured. We start with its millennia-old roots in *philosophy* and transition towards *linguistics, artificial intelligence, psychology* and culminate with contemporary *cognitive robots*. We challenge the dominant narrative outlined by Barsalou (2010) of a one-way influence of psychological notions of the *embodied mind* on robotics and artificial intelligence. Finally, we discuss our novel research programme - developing cognitive semantic models grounded directly in the visual world - a core thesis of the present dissertation.

2.3 Classical Theory of Cognition

A typical example of a classical theory of cognition is Alan Newell's (1980) *physical symbol systems* (PSS), which postulates that symbols are expressions with their origins in physical patterns and their manipulation gives rise to new expressions. Cognitive scientists who adopt the PSS, implicitly or explicitly, adhere to a so-called "hamburger model of cognition", consisting of *perception-cognition-action* modules, with a special focus on the intermediary *cognitive* stage, and treating *perception* and *action* as mere inputs and outputs (Willems & Francken, 2012). This standard view

has many core strengths, which are: (i) *type-token binding*, (ii) *inference*, (iii) *productivity*, (iv) *recursion*, (v) *propositions* and (vi) *scalability* (Barsalou et al., 2003, p.84). A core unifying perspective of much of cognitive science from these strengths is the *solipsistic* framework, which states that there is a definitive and non-negotiable boundary between the mind and the world. In this view, perceptions form the inputs from the world into the mind, while actions are outputs from the mind to the world. This idea also carries the implicit notion of reducing cognising agents to passive recipients of information from the world. The focus of traditional cognitive scientists on the “inner workings” of the mind, is in stark contrast to previous behaviourist paradigms studying *stimulus-response* pairs (for a historical perspective see Watson, 1913).

At the inaugural cognitive science conference, Allen Newell (1980) restated the fundamental contribution of both AI and computer science to the field of cognitive science through a detailed review of a *physical symbol system* (see figure 2.1). The apparent similarities between this view of cognition and the actual architecture of the modern computer are not a coincidence - Newell concluded his seminal article by stating that this insight into cognitive processing was “discovered indirectly while developing a technical instrument” (p. 182). In this review, we later argue that somewhat ironically, this is also the case for grounded cognition in two distinct ways. First, much of the philosophical and linguistic theory of grounding emerged as an oppositional framework to traditional computational methods. Second, early mechanistic formulations of grounding from AI and robotics predate the dominant theories of grounding from the 1980s and 90s.

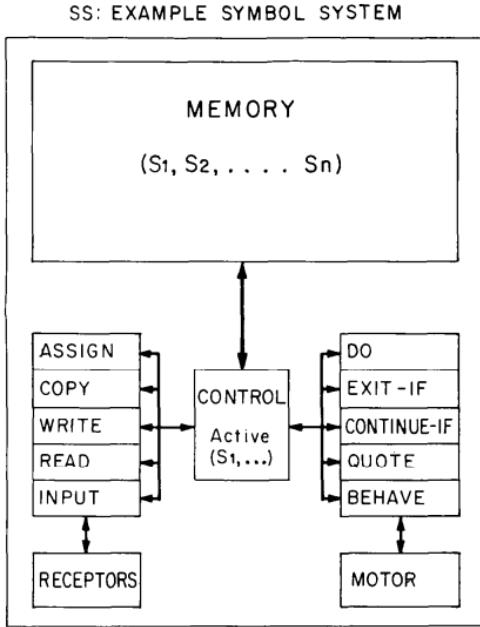


Figure 2.1: Example of a Physical Symbol System. Source: Newell (1980).

Glenberg (2015) postulates some key catalysts for the proliferation of the so-called “cognitivist perspective”, which overthrew the well-established tradition of studying *perception, action and cognition*. Two of the primary drivers, discussed by Glenberg are the emergence of *computer science* and the gradual realisation in linguistics that perceptual input alone could not satisfactorily account for the diversity of language. We will specifically focus on the first catalyst, as this is also widely used to delineate an allegedly insurmountable divide between the goals of computational modelling and grounded cognition. Newell and Simon’s (1976) *Physical Symbol System Hypothesis* claims that intelligent behaviours can arise in both humans and machines given that in both human and artificial intelligence, the core mechanism relies on the manipulation of abstract symbols. In some grounded cognition accounts (see Shapiro, 2011) this computational view is regarded as the antithesis to grounding.

Although some more modern cognitive architectures like ACT-R/E (Trafton et al., 2012) claim embodied properties due to the viability of sensorimotor re-enactments, we would still argue that these models are not grounded in the real world or realistic proxies such as naturalistic *sights* (e.g. photographs/videos) and *sounds* (e.g. audio files). The data inputs

feeding into the architecture typically consist of hand-coded features lacking any resemblances to the real world. The inputs are not grounded in real-world experiences.

2.4 Development of Grounded Cognition

Grounded theories of cognition (e.g. Barsalou, 1999; Lakoff & Johnson, 2008) challenge traditional views of cognition by claiming that conceptual representations are grounded in sensorimotor experiences, and also processed at this level, and not in an abstract, amodal manner (*see figure 2.2*). A variety of related but different terms such as *embodied cognition*, *situated cognition* and *enactive cognition* refer to positions against classical theories of cognition. Grounded cognition is a unifying umbrella concept spanning all of these more specialised definitions. Grounded theories initially gained traction within the humanities, in particular, *cognitive linguistics* and *semiotics* (Lakoff, 1993; Talmy, 1996). In contrast, disembodied cognitive theories generally favour a system in which conceptual representations are symbolic and abstract, and therefore qualitatively different from sensory and motor codes. Fodor's (1983) modularity hypothesis is an example of such a perspective since it states that central cognition is amodal but that efferent and afferent connections to the world are modal.

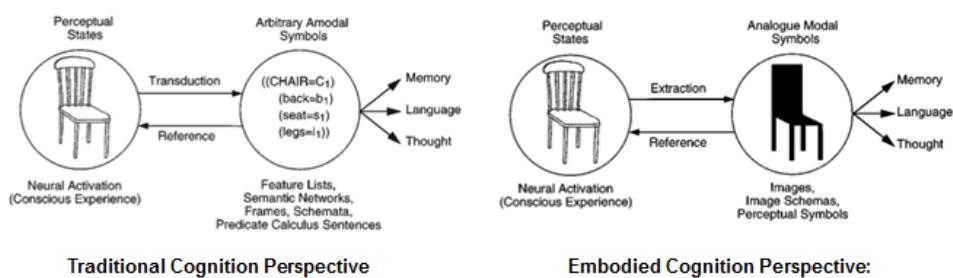


Figure 2.2: Comparison of traditional and embodied symbol states. Source: Barsalou (1999).

Barsalou (2010) reviewed the past, present and future of grounded cognition spanning *philosophy*, *linguistics*, *psychology*, *neuroscience* and only included two passing citations on *behaviour-based robotics* and *artificial intelligence*. However, according to Barsalou (2010), the future of grounded cognition is closely tied with physical robots, because these fields provide an ideal test bed for grounded cognition theories. Somewhat surprisingly, the review fails to outline any contributions of the field of robotics as a historical antecedent to the rise and dominance of grounded cognition and its influence in psychology over the past thirty years. According to Barsalou's (2010) timeline, grounded cognition is rooted in *philosophy* and *linguistics*, while future developments are likely to come from *cognitive ecology*, *neuroscience* and *robotics*. Although we broadly agree with this view, we claim that an overlooked aspect of the history of grounded cognition is artificial intelligence and robotics. When discussing the origins of grounded cognition, Barsalou's (2010) review merely states the following: "[I]n robotics, Brooks (1991b) and Kirsh (1991) advocated incorporating the environment and the body into a new generation of robots." (p. 718).

Similarly, in a related review article by Mainzer (2009) titled "*From embodied mind to embodied robotics: Humanities and system theoretical aspects*", the same argument is made about linguistic, semiotic and philosophical ideas shaping contemporary embodied robotics. Surprisingly, Mainzer does not include a single citation of Brooks-style robotics and AI influencing embodied theorising. Although traditional symbolic computational references are made, like Tarski's (1935) *correspondence theory of truth*, outlining the *isomorphism* between the real world and symbolic abstractions. There is current literature detailing the interrelated origins of grounded cognition and ecological psychology (see Clark, 2008). Although to the best of our knowledge, there are no such equivalent reviews for the fields of AI and robotics. In this introduction, following an overview of some of the fundamental *philosophical*, *linguistic* and *empirical* milestones, we outline pioneering research from the fields of *artificial*

intelligence and robotics between the years 1971 and 1991, which helped shape the rise of mechanistic formulations of grounded cognition in psychology. The research outlined, on grounded cognition in AI and robotics, either pre-dates or concurrently develops with other philosophical and linguistic research programmes between 1980-1999.

Grounded cognition has had a great many research outputs in *cognitive linguistics* (Johnson & Malgady, 1979; Johnson, 1981; Johnson & Henley, 1988; Talmy, 1996; Bundgaard & Østergaard, 2007), but has since influenced research on *memory* (Roediger, 1980), *emotions* (Niedenthal et al., 2001, 2005; Markman & Brendl, 2005), *action understanding* (Tucker & Ellis, 1998), *judgement and decision making* (Schnall et al., 2008), and even applied fields such as *organisational psychology* (Harquail & King, 2010), *educational psychology* (Sousa, 2010), *clinical psychology* (Brown & Ryan, 2003) and *cognitive neuroscience* (Matheson & Barsalou, 2017).

2.4.1 Grounded Cognition: Roots in Philosophy and Linguistics

In philosophy, Merleau-Ponty (1945) emphasised the close relationship between perception and action and theorised the basis of conceptual representations originating from grounded interactions in the real world (Anderson, 2003). However, the most impactful criticisms to the physical symbol system perspective of Newell and Simon (1976), as well as other related hypotheses (e.g. Atkinson & Shiffrin, 1968) came from a thought experiment by Searle (1980). John Searle's *Gedankenexperiment* revolves around imaging a scenario whereby a non-Chinese speaker finds themselves in a sealed room with two slots, one from which they receive Chinese logograms, mostly a series of "meaningless squiggles" to the non-Chinese speaker (p. 3) and another slot from which they can output perfectly grammatical phrases of Chinese symbols. The only aid the non-Chinese speaker inside the locked room has is an extensive rulebook mapping inputs of Chinese logograms with grammatically correct outputs of Chinese logogram sequences. Searle asks: *does the person inside the room*

speak and understand Chinese? The person inside the room is performing operations on formally specified input-output mappings, as per the rulebook, but are they fluent in Chinese? Searle answers this question negatively, and most critically, parallels this to the physical symbol system, a closed system of abstract amodal symbols (devoid of perception and action) from which meaning of symbols only stems from its interrelations with other abstract amodal symbols. Similarly, Harnad's (1990) "symbol merry-go-round" argument succinctly summarises why a closed system of abstract symbols cannot be the underlying theory of meaning, given this circularity and that only by breaking away from the vicious cycle by grounding those symbols in *perception, action* and other bodily states such as *emotion*, does meaning genuinely emerge.

At first glance, the relatively recent hype and interest in grounded cognition can seem like an entirely new "paradigm" discovered in the late 20th century. However, in reality, it is more of a resurrection of ancient philosophical and even religious ideologies, a gradual elaboration of philosophical and metaphysical principles through the deployment of original and elegant research paradigms. The origins of grounded cognition are typically traced back to the philosophical ideas of the ancient Greek philosopher Epicurus (341 - 270 BCE), based on his atomic materialist views of matter being at the heart of everything - from physical to mental objects, and that genuinely disembodied or symbolic states disconnected from external reality are impossible (Barsalou, 2008). Even abstract concepts such as *good* or *bad* are omitted from explanatory descriptions of human thought; instead, the focus is on actual physical instantiations like *pleasure* and *pain* (Shapiro, 2011). Numerous grounded cognition theorists (e.g. Clark, 2008; Barsalou, 1999; Glenberg, 2015) have shed light on the historical pedigree of the idea that higher-level cognition is not independent of perception and action, one of the core pillars of grounding in cognitive science and artificial intelligence. This scholastic lineage can also be dated as far back as Aristotle (367 - 347 BCE) and the notion of Aristotelian metaphysics of emotions

(Gabbe, 2016). Grounded philosophies of cognition even provide coherent phenomenological accounts of the *embodied soul*, which according to Wright (1991) was used in the 17th century for medical diagnoses.

These Western philosophical origins have been extensively discussed in the grounded cognition literature (see Haugeland, 1993; Barsalou, 2008; Mainzer, 2009), although the ancient Indian thought system *Samkhya-Yoga* has been entirely overlooked. The present author proposes that this South Asian history is more directly relevant to grounded cognition, especially given the increasingly popular intertwined nature with modern-day Buddhist philosophy, mindfulness and grounded cognition (e.g. Watts, 2013). Samkhya, also written as Sankhya, is one of the core pillars of Indian philosophy, originating circa 5th century BCE, which has strong rationalist and reductionist underpinnings (Sinha, 1979) and is widely acknowledged to be strongly dualist, given its distinction between *buddhi* ("intellect") and *ahankāra* ("consciousness"). However, when Samkhya is combined with Yoga, to form Samkhya-Yoga, the focus is on applied knowledge, consisting of integrating the three "modules": (i) *perception*, (ii) *inference* and (iii) *memory*. This integration leads to *abstract knowledge*, translated from *Purusha*, meaning 'pure consciousness'.

Furthermore, it also seems that Samkhya-Yoga, as interpreted through Sinha's (1979) translations, advocates that all knowledge be grounded in sensory inputs, more specifically the visual and auditory modalities. Our delineation of the philosophical origins of grounded cognition indicates that early philosophical/religious texts sought to find common relationships between the mind, body and environment, which share some surface-level similarities with psychological theories of grounding. More recently in Europe, during the *age of reason*, enlightenment philosophers such as Immanuel Kant (1724 - 1804) also claimed that all mental sensations are affected either directly or indirectly through physical objects (Svare, 2006).

At the intersection of philosophy and linguistics, resides Lakoff and Johnson's (1980) highly influential *conceptual metaphor theory* (CMT), with the core trilogy¹ of publications cited approximately 100,000 times. To contextualise this, only the top three *most highly cited* publications out of the top 100 reported by *Nature News* (Van Noorden, Maher, & Nuzzo, 2014), contain more lifetime citations than received for CMT. In CMT, concepts are represented by relation to another more basic domain. Lakoff and Johnson develop a diagrammatic framework, known as *conceptual blending*, for explaining the process through which typically more abstract conceptual domains (known as the *target domain*) are metaphorically grounded in more basic conceptual domains (called the *source domain*). For example, *happiness* (the abstract target domain) can be interpreted via the metaphoric extension of *verticality* (the source domain), and lead to people generating and understanding linguistic utterances such as "I am feeling a bit down today". Another example of a target-source relationship would be difficulty-heaviness. Although exceptions exist, the typical pattern is such that the source domain is grounded in sensorimotor content, while the target domain consists of abstractions or reactions to the sensory input. We argue that conceptual blending is a descriptive framework lacking clear specifications or predictions. However, Lakoff and Johnson's original theory is philosophical and requires a robust empirical basis. Also, Turner (1996) and Gibbs, Ellison and Heino (2006) critique that such metaphoric extensions are not only a literary artefact but also associated with human cognition (Barsalou, 2008).

In their more recent work titled *Philosophy in the Flesh*, Lakoff and Johnson (1999) fail to make any effort of relating their underlying ideas to proposals in AI and cognitive science and furthermore omit to appreciate non-philosophical approaches to grounding. Therefore, unfortunately,

¹ Lakoff and Johnson (2008), Lakoff (2008) and Lakoff and Johnson (1999) being respectively cited 55,564, 27,413 and 16,390 times.

their philosophical work is challenging to evaluate from both an empirical and a computational perspective. Even Gallese and Lakoff's (2005) descriptive model of bootstrapping sensorimotor information for conceptual processing falls short of being detailed enough to constitute a formal theory. The descriptive model suggests that concepts are fundamental building blocks from which linguistic meaning and reasoning emerges based on the *neural theory of language* (NTL) and *theory of cogs* (ToC). In Feldman and Narayanan's (2004) NTL, the same neural structures implemented for perception and action in the real world are also used offline for abstract reasoning through inference mechanisms based on ToC's use of generic X-schemas that allow for abstract logical inferences based on a symbolic inference-engine architecture outlined in Narayanan (1997). We find that, like Barsalou (2010) and Mainzer (2009), Lakoff and Johnson (1999) and Gallese and Lakoff's (2005) also fail to credit the fields of artificial intelligence and robotics contributing to the rise of grounded cognition.

2.4.2 Grounded Cognition: Roots in Psychology

In cognitive science, the origins of grounding cognition typically point toward Gibson's (1979) *theory of situated action*, focusing on the interplay between close action-perception coupling dynamics leading to an emergence of complex behaviours, in the absence of symbolic modal or amodal representations. Gibson's theory of situated action, as well as his related work on the brain's role in vision strongly deny the solipsistic framework adopted by mainstream cognitive scientists. However, stalwarts of traditional symbolic cognitive science (e.g. Fodor & Pylyshyn, 1981) went to great lengths at discrediting Gibson's views as neobehaviourism with different terminology (Shapiro, 2011).

Garbarini and Adenzato (2004) suggest that grounded cognition stems from the confluence of James Gibson's *theory of affordances* (Gibson, 1979) and Eleanor Rosch's *principles of categorisation* (Rosch, 1977). The first

core principle outlined by Rosch is *cognitive economy* - the effort to distinguish between two objects must be proportional to the advantages of this distinction and perceived world structure. Second, perceptual information reaches us in a *non-arbitrary* or *meaningful* way, for example, focusing on the role of functional needs of a particular object being viewed in different ways by different organisms.

There is an abundance of empirical demonstration experiments showing the embodied or grounded nature of cognitive processing (see Barsalou, 2010). Here, we will limit our focus to only semantics, which has gained a great deal of attention. Much of this evidence relies on the premise that if semantic knowledge is rooted in low-level sensorimotor areas in the brain, then these should also be active during retrieval of higher-order semantic knowledge (Mahon & Caramazza, 2008). For example, in an fMRI study, Hauk (2004) found that reading action words (e.g. *pick*, *lick*, *kick*) activates premotor and frontal areas associated with performance of actions with corresponding body parts (e.g. *hand*, *mouth*, *leg*), which Hauk interprets as supporting embodiment. This link between embodied accounts and neuroscientific evidence is a recurring theme within the grounded cognition literature, not least because the seminal paper outlining *perceptual symbol system* (Barsalou, 1999) emphasises the natural fit between embodiment and brain mechanisms, based on simulation theory's reliance on modality-specific activations, and the brains modality-specific areas.

2.4.3 Grounded Cognition: Roots in AI and Robotics

In the domain of *artificial intelligence* and *robotics*, mechanistic frameworks emerged in the early 1970s, which closely resembled many aspects of *ecological psychology*. Most of these historical frameworks are not related to grounding meaning per se but are more focused on practical navigation tasks along with the computational challenges of planning and executing motor sequencing in increasingly complex and realistic

environments. Munson (1971, p.1) was perhaps one of the first roboticists to acknowledge that a fully formalised model of any realistically complex environment was unfeasible due to a range of reasons such as (i) *lack of sensory precision*, (ii) *lack of complete descriptions*, (iii) *lack of effector (motor) precision* and (iv) *memory storage limitations*. Munson (1971) developed one of the first formalised frameworks for the *planning*, *execution* and *monitoring* of robotic systems in an uncertain environment. The general formulation was inspired by the inadequacies of problem solvers available at the time to cope with real-world dynamic environments when only using first-order predicate calculus. Munson's (1971) work cited the problem that arose in "robotry" as a result of the disconnect between an agent's internal and external environments and their links to acting and planning. In many ways, these days, one might be tempted to categorise Munson's (1971) AI research framework as GOFAI and unrelated to grounded cognition, Brooks-style AI or present-day cognitive robotics. However, there are many reasons why, despite the implementation-level similarities with GOFAI (decision trees and maximising utility), the approach is a precursor to grounded cognition.

First, Munson recognises the additional problems posed through the distinction between internal and external states. Second, in response to the first dilemma, there is an emphasis on relating *planning* (cognition), *execution* (sensorimotor outputs) and *monitoring* (sensorimotor inputs) as opposed to conceptualising each stage as a distinct process like in the classic "sandwich model" of cognition. Third, there is an explicit acceptance of the limitations of *game-theoretic* formulations of decision making under uncertainty due to the inherent complexity of the environment in spite of retaining the concept of utility for comparisons with the "ultimate rational" execution (p.9). Fourth, the robot's knowledge space is dynamically defined using a three-element tuple, $\{m_1; P; C\}$, which we interpret as a *grounded intelligence structure*, whereby *model relations* (m_1), *state of the plan* (P) and *control information* (C) are entangled. These four points taken together

suggest a tight coupling between *perception* and *action* - the fundamental tenet underlying grounded cognition.

Furthermore, Munson's combination of a dynamically-coupled knowledge space for monitoring, planning and executing motor commands makes a strong case for the first-ever adaptive AI platform. Therefore, we argue that even though Munson never uses the term "grounding" or "behaviour-based", the highly original ideas expressed and formalised are strong precursors to dynamic perspectives of grounded cognition as well as Brooks-style behaviour-based robotics. In the words of Brooks (1999, p.8), "the coupling of perception and actuation systems...is the cornerstone of behaviour-based robotics".

We mentioned that most examples of early AI and robotics research were not associated with semantics, although there is one notable exception. One of the first technical breakthroughs in a natural language-based grounding of artificial intelligence originated from John F. Sowa in the mid-70s, who at the time was an IBM researcher. Sowa (1976) developed a novel type of knowledge structure, called *conceptual graphs*, which was a significant step forward in both AI and grounding meaning because it allowed for a seamless interface between users and a database interface. The system allowed users to access database items without knowledge of the technical details underlying the database or metadata structures but via the conceptual graph abstraction. Additionally, the conceptual graphs permitted for relations to be generated between concepts to encode functional dependencies. This approach developed at a time when large volumes of physical (usually paper) records were transcribed into large computational databases. Conceptual graphs provided a formal notation for human users to interact with the conceptual graphs by translating a natural language query into a conceptual graph, which in turn would be converted into a machine-readable query to access the most relevant information from the database. To the best of our knowledge, this was the first prototype of a simple, yet functional, natural language-based search

functionality. A further extension and elaboration of some of the original features of conceptual graphs in Sowa (1979) outlined the unique property of not only being an interface but also encoding, storing and decoding semantic associations from the real world. Sowa's (1976, 1979) conceptual graphs also relied on Heidorn's (1975) NLP parser and therefore could be grounded in the real-world linguistic environment. This is an early example of grounding semantics in real-world language.

In a series of technical and philosophical publications, Rodney Brookes developed a new research programme for AI, called *behaviour-based robotics* or *Nouvelle AI*, which went beyond some of the pioneering early efforts of Munson (1971) and Sowa (1979). Behaviour-based AI moved away from the "sandwich model" of *perception-cognition-action* by allowing *cognition* to be an emergent property of the dynamic interactions with the world through perception and action. Brooks' underlying philosophy was that robots should operate in the types of environments humans operate in, different from the sterile lab conditions under which earlier versions of GOFAI-based robots like the *Shakey* operated (Brooks, 1999).

In the first landmark publication introducing the ideas underlying behaviour-based robotics, although not the term itself, Brooks (1986) presented the *layered control system* for mobile robots. Traditional robots typically consisted of a wide range of modules and sub-modules for highly specialised tasks. However, in Brooks' architecture, the modules were structured in hierarchical layers, with increasing levels of competence as one moved from the more basic to the higher-order layers. Critically, higher layers can entirely *subsume* the functions of the lower-level layers, even though the addition of higher layers does not impact the lower levels' functions. Thus, *sensors* directly send inputs to all the hierarchical structured modules, which in turn, all send outputs to the *actuators*. The different levels have different functional properties determined by the make-up of the individual layers comprised of multiple modules (or processors) that only send messages to each other within the same layer. A

single module is an augmented FSM (*finite state machine*) with the ability to read-write LISP data structures. However, these modules only have the “shortest range” memories because their single element buffer can be overwritten with new information before it is read by another module within the same layer (Brooks, 1986). In this initial publication, there were three layers. The *zeroth level* ensured that the robot did not collide with other objects. The *zeroth level + first level* enabled the robot to move around aimlessly without crashing into other objects. Lastly, the *zeroth level + first level + second level* allowed the robot to demonstrate more exploratory behaviours such as moving towards corridors with free space. From the robot moving around successfully using the behaviour-based subsumption architecture, Brooks (1986) ruled out the need for a central processor.

Brooks (1989) managed to use the behaviour-based subsumption architecture to control a six-legged robot, which has significantly harder control system challenges that need to be overcome (Brooks, 1999). This insect-like robot was called *Genghis*, one of Brooks’ more famous creations, now located in the *Smithsonian Air and Space Museum*. Genghis’ AI architecture was built incrementally using a network of 57 augmented finite state machines (modules) structured hierarchically to enable Genghis to perform a range of increasingly complex behaviours from *standing up* and *simple walking* to the highest level, *steered prowling* which allowed Genghis to follow slow moving objects or people. Such behaviours demonstrated that highly sophisticated and co-ordinated macro-level behaviours could emerge from a broad set of layered micro-level primitive functionalities. In other words, fundamental building blocks combined in the *right way* could lead to the emergence of complex adaptive behaviours.

Despite the successful outcomes of Brooks (1986, 1989), both of them were still based on *in situ* coupling of the sensory and motor systems, without manipulating internal representations, and hence, lacked demonstration of more sophisticated behaviours. This shortcoming was also one of the reasons for the behaviour-based approach to AI receiving

strong criticisms from more traditional AI and robotics researchers. As a direct response to this mounting criticism, Mataric and Brooks (1990) repurposed the subsumption architecture in a new robot called *Toto*, to demonstrate the feasibility of acquiring distributed map representations grounded in navigation behaviours. Moreover, these maps were dynamically deployable in real-time navigation. This dynamic behaviour was the first example of complex mapping representations being encoded directly from grounded interactions. Once more, *Toto* was based on the layered subsumption architecture and therefore contained no central processor, and there was overlap between the modules representing *Toto*'s spatial position and planning the actions. Like with Genghis, *Toto* could do a range of increasingly more complex behaviours as a result of the higher-level layer interacting with the lower layers.

The most significant difference with *Toto* was that unlike in previous subsumption architectures, landmark identification automatically led to the relative mapping of the different landmarks. However, even this process was not centralised. Maja Mataric, then a graduate student in Rodney Brooks' lab, demonstrated the feasibility of graph-like structures being encoded in a decentralised information system comprised of a series of augmented finite state machines (Mataric, 1990). Each finite state machine is a node in the map-style representation. Mataric and Brooks (1990) utilised this distributed mapping technique to allow *Toto* to acquire a map of its environment through gradual exploration. The so-called map-nodes, which are empty at compile time, gradually start obtaining information (the type of landmark and corresponding direction) in parallel in a distributed manner. *Toto*'s localisation is performed by all the nodes comparing their details with the current information, to activate the corresponding node. However, if a matching node is not found, this is interpreted as a new landmark and is added to the graph. Finally, nodes spread their activation to other nodes in the direction of travel, which leads to more sophisticated planning repertoires (Mataric & Brooks, 1990). As one

would expect, despite the initial starting position of Toto varying across trials, over time, landmarks clustered together based on their spatial layout. The robot Toto was also capable of dynamically changing its goals stored in *long-term memory* and behaving in a manner consistent with flexible plans being generated from the distributed map representation. This early study demonstrates the incremental acquisition, representation and manipulation of a distributed graph-like representation through the close coupling of sensorimotor and environmental signals and without the need for a central processor. Unsurprisingly, in a later review of this work, Brooks (1999, p.37) claimed that “[he] view[s] this work as the nail in the coffin of traditional representationalism”. However, as we shall discuss in *section 2.5* of this chapter, Brooks (1999) ended up adopting a more balanced conclusion and even accepting that behaviour-based AI failed to live up to the original expectations. We also suggest that Mataric and Brooks’ (1990) robot Toto is a physical instantiation of the frameworks and models put forward by earlier grounded robotic principles from Munson (1971) and Sowa’s (1976, 1979) work on demonstrating the feasibility of using conceptual graphs grounded in natural language to represent meaning.

Brooks (1991c) summarises the successes of the behaviour-based robotics paradigm initiated through the development of the subsumption architecture in 1986, and subsequent refinements and demonstrations through robots like Genghis and Toto. Critically, we argue, Brooks (1991c) highlights two central themes, which are sufficiently significant not to be hidden in the main body of the article but highlighted with bullet points on the first page of the article. We include the excerpt below from Brooks (1991c, p.1227):

- “*Situatedness*: The robots are situated in the world—they do not deal with abstract descriptions, but with the “here” and “now” of the environment that directly influences the behavior of the system.

- *Embodiment*: The robots have bodies and experience the world directly — their actions are part of a dynamic with the world, and the actions have immediate feedback on the robots' own sensations.”

Brooks' (1991c) quotes on *situatedness* and *embodiment* are eight years before Barsalou's (1999) influential *perceptual symbol hypothesis* provided a descriptive argument in favour of tighter couplings between sensorimotor activities and high-order cognition. Also, Brooks' (1991c) theoretical overview is based on more than five years of robotics experimentation demonstrating the feasibility and benefits (e.g. increased efficiency and speed) of grounding higher-order intelligent operations such as planning using situated graph representations in an unpredictable environment. Therefore, the shift from a modular cognitivist *theory of mind* or traditional GOFAI to grounded or dynamic cognitive representations not only preceded the rise of grounded cognition in psychology (Barsalou, 1993, 1999) but was founded on rigorous mechanistic models. Furthermore, Brooks-style grounding in robotics, because of the nature of the discipline, provided functional models of non-separation of data and computation given the distributed representation of both across the same modules.

2.4.4 Summary

In conclusion, unlike Barsalou (2010), we argue that the aforementioned developments in AI and robotics (Sowa, 1976, 1979; Munson, 1971) in conjunction with the five seminal theoretical papers from Rodney Brooks, published between 1985 and 1991, support our claim that grounded cognition's origins are also partly rooted in AI and robotics. Naturally, as discussed, we also acknowledge ancient Western and Asian philosophical influences as well as from linguistics more recently. Theories of grounding in cognitive psychology only emerged in the mid-to-late 1990s (Barsalou, 1993, 1999). Although, lesion studies from *cognitive neuropsychology* did demonstrate a role of the modal systems in

representing category knowledge (Warrington & Shallice, 1984). However, it was not until Barsalou (1993), that there was a specific descriptive psychological theory of human knowledge being grounded in our sensorimotor modalities. The same is also the case for some linguistic conceptions of grounding (e.g. Lakoff, 1993). Similarly, Varela et al. (1991) book titled *The Embodied Mind* emerged after the core behaviour-based robotics developments. Varela et al. provide a detailed overview of Brooks' robotics experiments. The cognitive descriptions and theories of the human embodied mind and the importance of being embedded in real-world dynamics only appeared several years following the aforementioned conceptual graphs grounded in language (Sowa, 1976) or behaviour-based robotics (e.g. Munson, 1971; Brooks, 1986). Additionally, a range of philosophical articles fleshing out the importance of behaviour-based robotics from the *embodiment* and *situatedness* principles (Brooks, 1990, 1991a, 1991b, 1991c) also pre-date the earliest publications from cognitive psychology. Grounded cognition's origins are, therefore, based on AI and robotics.

2.5 Limitations of Grounded Artificial Intelligence

Contemporary efforts of behaviour-based robotics are broadly still based on many of the core principles outlined in Brooks (1991c, 1999). More recently, a plethora of interrelated research domains have branched out from the original behaviour-based robotics research of the late 1980s and early 1990s, such as *developmental robotics* (Asada et al., 2001; Weng, 2004), *cognitive robotics* and *evolutionary robotics* (Minato & Ishiguro, 2007; Nolfi et al., 2016), which have mainly advanced from an *engineering* and *computer science* perspective. In other cases, so-called *embodied-roboticists* use tried-and-tested analytical methods, such as dynamical systems modelling using differential equations, to generate perception-action couplings (e.g. Pfeifer & Scheier, 2001). These types of approaches are reminiscent of the

pioneering research of Munson (1971). However, these key developments are not entirely relevant for our present purposes of investigating *natural human cognition* mechanistically.

Mainzer (2009) praises the behaviour-based robotics paradigm for being superior at accounting for the dynamic emergence of new robotic behaviours from simple modules running in hierachal unison between sensations and actions. Still, Mainzer (2009) also critiques grounded roboticists for merely assuming that macro-level *self-organising* principles automatically lead to the desired goal of human-like intelligence (p. 296). Therefore, the suggestion is for roboticists and AI researchers to understand the processes of controlled emergence better to build a system that might gradually display more human-like behaviours if the artificial system is guided appropriately.

Barsalou (2010) claims that grounded theories will eventually synthesise with classic symbolic and sub-symbolic perspectives and the new developments in behaviour-based/embodied robotics will allow for mechanistically evaluating different embodied theories. We agree, that there has been an increased effort to investigate the mechanistic nature of embodiment and higher-order cognition (e.g. Cangelosi & Riga, 2006; Pezzulo et al., 2011), but this has primarily focused on ensuring embodied constraints are operationalised through the use of hardware (e.g. physical or virtually embodied robots). We believe much of embodied robotics merely consists of ensuring that physical or virtual effectors execute specific actions. There is, however, a relative lack of research on grounding higher-order cognition in the environment, as opposed to directly embodying abstract concepts like numbers in lower-level sensorimotor contingencies like robotic finger configurations (e.g. De La Cruz et al., 2014).

An overview of behaviour-based robotics would not be complete without an outline of some of the core limitations. We first highlight some of the dominant criticisms from Dreyfus (2007), followed by our suggestions. Although Hubert Dreyfus is well-known as an “ideological

foe” of AI (Dennett, 2006, p.434), most of the philosophical points raised more than a decade ago are still relevant and largely remain unanswered by the AI community. Many of Dreyfus’ critiques of grounded AI models, especially behaviour-based ones, have somewhat inadvertently been addressed in this dissertation’s later computational and empirical research chapters. We, therefore, include the labels (C1 to C5) to highlight some of the critical challenges posed by Dreyfus, which we address in the present dissertation.

Dreyfus (1997) raises one of the most robust sets of criticisms not only against behaviour-based but AIs/robots in general. The main criticism is that computational systems are inherently unable to account for an *intentional agency*, which Froese and Ziemke (2009) interpret as computational intelligence not having an “*understanding* of their situation” (p. 467). In other words, Dreyfus critiques the inability of AI agents to acknowledge relevant features under real-world conditions based on the system determining these particular features to be salient as opposed to being pre-determined by a human modeller (Froese & Ziemke, 2009). We strongly agree that historical as well as current computational modelling research aiming to mimic human intelligence are unable to demonstrate sensitivity to *context-dependent relevancies* (C1). However, this critique has a corollary criticism, namely that most computational models, grounded or otherwise, do not focus on *understanding* (C2). We interpret the term “*understanding*” as humans’ (and other animals’) ability to navigate their surroundings full of interdependent conceptual associations. It is difficult to imagine a robot genuinely understanding *anything* if it does not have human-like semantic representations in the first place.

The third criticism is that there is a sense of “inadequacy of current embodied AI for advancing our scientific understanding of natural cognition” (Froese & Ziemke, 2009, p.467). In our view, the inability for computationally grounded models to provide sufficient insights into human cognition (C3) is directly related to the tried-and-test GOFAI

practice of modellers determining a phenomenon of interest and then *hand-coding* so-called “grounded representations” (e.g. Hoffman et al., 2018). This leads to a certain degree of circularity in the modelling processes, mainly if models are evaluated by behaviours displayed. Another criticism of behaviour-based robots is that embodiment does not overcome the grounding problem (Froese & Ziemke, 2009, p.467). We argue that this dilemma of being embodied but not necessarily grounded in the environment (C4) originates from an over-emphasis on either physical or virtual effector-systems at the expense of grounding the robot in the messiness and dynamics of real-world changes. In our view, grounding a computational model or robot with the objective of having human-like experiences is unrealistic if we fail to provide a sufficiently realistic environment. For example, for a cognitive model, if we want to input visual object features, we should not resort to binary feature lists (e.g. 10011000) when instead we could use naturalistic photographs.

Lastly, but quite critically, Dreyfus’s (2007) fifth major criticism of embodied robotics is the *frame problem* (C5). Dennett (2006, p. 434) eloquently phrases the severity of the frame problem with “Hubert Dreyfus and John Searle are tempted to compose obituaries for the field [AI], citing the *frame problem* as the cause of death”. Originally, the frame problem was narrowly formulated by McCarthy and Hayes (1969) to describe the issue of using *first-order logic* (FOL) systems to generate sufficiently exhaustive and useful axioms that can describe the properties and interrelations of a robot’s environment. However, Dennett’s (2005, 2006) broader definition of the frame problem describes the inability of AI systems to comprehend the most relevant aspects of a given scene autonomously, without the human-modeller hand-coding the relevant features *a priori*. In a given scenario, how could an AI agent solve a particular challenge by automatically *only* focusing on the most important features and associated actions for a given task? This “looser philosophical formulation” of the *frame problem*, as opposed to the original FOL-based one, is intertwined with what most AI

researchers associate with the *relevance problem* - autonomously *understanding* critical aspects of the environment that are relevant.

Neither traditional nor behaviour-based AI approaches have sufficiently addressed any of the above five challenges. We hypothesise this failure is a result of not focusing on automatically extracting semantic associations from the real world - the key ingredient for context- and task-specific *understanding*. As AI researchers, both GOFAI and behaviour-based advocates, increasingly focused on narrower problems, typically with specific commercial applications, the challenges mentioned above actually became harder to address. Unsurprisingly, like in GOFAI, one of the main unacknowledged shortcomings of the behaviour-based robotics paradigm also consists of (i) an excessive focus on dynamic planning and action in spatial domains or well-defined games and (ii) a lack of extracting complex meaning structures from real-world sources. This is a gap in the extant literature we would like to address in our current research programme. For far too long have mechanistic theories of intelligence, either from cognitive modelling or AI and robotics, ignored real-world *meaning*. However, at least for higher primates, it seems realistic to assume that much of our intelligence originates from our semantic memory system. Acknowledging this might also help address the five challenges defined in Dreyfus (2007).

2.6 Towards a Computational Grounded Semantics

Intelligence is a broad and multifaceted phenomenon, with a great deal of debate on its origins, structure, development and manifestations across a range of species. Recently, there have even been proposals for accepting that plants display *minimal cognition*, from environmentally adaptive behaviours. This has led to the term *plant cognition* emerging directly as a result of the increasingly widespread influence of grounded cognition in theories of human intelligence (Garzón & Keijzer, 2011). In psychology, cognition is associated with, for example, *perception*, *learning*

and memory and judgement and decision-making processes. Despite cognitive psychology's plethora of interrelated fields of study, we suggest that *semantics* is likely to be the main *component* for intelligence to emerge in both cognitive science and artificial intelligence. Semantics is the study of meaning-making and representation through our interconnected experiences. Within the domain of semantics in cognitive psychology, *concepts* constitute the basic building blocks of meaning. Our emphasis on *cognitive psychology* sketches the focus of this dissertation. We omit discussions of *formal semantics*, *logic* and *semiotics*, as that is beyond the scope of our present research programme. Laurence and Margolis (1999, p.3) argue that concepts not only are the most fundamental constructs in *theories of mind* but also that this leads to plenty of controversies and unsettled debates on the very nature of semantics itself. In our view, the importance of grounded cognition in cognitive science and artificial intelligence will primarily be based on the relative merits of this more ecological stance to account for human-level semantic processing.

In traditional cognitive science, a widely accepted theory of semantic cognition is founded on *spreading activation* models. It is worth pointing out that we view these models, despite their historical pedigree in GOFAI, as sub-symbolic models, and very much compatible with some forms of embodied and grounded modelling, even though some proponents of grounded cognition (e.g. Chemero, 2011) overstate the differences between such spreading activation models and grounded cognition principles.

Collins and Loftus (1975) created the first spreading activation model of semantic cognition, by extending the work on earlier static network models (Quillian, 1967). These models generally contain five core features. First, the graph-theoretic concepts of nodes and links are used to represent concepts and their relations respectively. Second, concepts acquire their meaning through their correlations with other concepts. Third, memory retrieval from these networks is equivalent to activating the

internal representation. Fourth, activation propagates from one concept to other concepts via the links. Fifth, concepts can typically build up residual activation which increases their likelihood of being activated in subsequent retrievals. In the classic model of semantic processing by Collins and Loftus (1975), in addition to having a semantic network based on the five aforementioned principles, there is a secondary lexical network, which is organised based on phonetic similarities (e.g. *car* and *bar* would be connected although *car* and *bus* would not be), which is linked to the semantic network. The activations across different nodes are summed (weighted by the activation strength) with activations spreading greater distances taking longer to spread (McNamara, 2005). Although these models have been highly influential in cognitive science (Neely, 1977; Crestani, 1997; Anderson, 2013), they have received little attention within the grounded cognition literature. This illustrates one of the critical challenges of the GC literature, in that it consistently attempts to distance itself from “traditional cognitive science” without specific references to influential models that have been empirically tested for over four decades. In our research programme, we aim to build bridges between historically disparate fields in pursuit of grounding meaning in everyday surroundings.

A fundamental philosophy spanning this present dissertation is the notion of “understanding by building” (Froese & Ziemke, 2009, p.468). However, we argue that contemporary computational attempts of grounding semantics are in a similar state of GOFAI, before the era of behaviour-based robotics revolutionised the field by transcending idealised and static grid worlds. Contemporary grounded models are exceedingly pre-occupied with sensorimotor interactions of physical hardware as opposed to naturalistic environments. Currently, cognitive modellers are unable to automatically or semi-automatically ground meaning without recourse to hand-coding features. Therefore, we aim to address this shortcoming by *extracting, processing, transforming and representing* complex

semantic associations grounded in the visual disorderliness of the real world, which comes naturally to humans and other animals but remains a challenge for computers. We believe this will further a mechanistic understanding of grounded cognition.

2.7 Summary

In contemporary cognitive science, there is a dominant narrative (Mainzer, 2009; Barsalou, 2010) of grounded cognition's roots stemming from ancient Western philosophical ideas and, more recently, linguistics. Furthermore, these reviews also suggest that grounded cognition is likely to see promising future outcomes from the fields artificial intelligence and cognitive robotics. We only partially agree with this. In this chapter, we have provided a novel account of an alternative historical influence - the overlooked Indian philosophical and religious influences of *Samkhya-Yoga*, which pre-dates the Western philosophical viewpoints of grounding and embodiment by almost 200 years. More importantly, we also demonstrate that influences from artificial intelligence and behaviour-based robotics also predate (e.g. Munson, 1971; Sowa, 1976; Brooks; 1986) the earliest literature on embodiment in cognitive psychology (e.g. Barsalou, 1993, 1999). Support for this interpretation of our timeline is strengthened by both explicit theorisings of *situatedness* and *embodiment*, in Brooks' (1991c) review article aimed at a general scientific audience and Varela et al.'s (1991) detailed overview of Brooks' behaviour-based robotics paradigm. In our literature review, we outline some of the hidden influences of robotics and artificial intelligence but also agree with Barsalou (1993) that these fields will be necessary for evaluating theories of grounded cognition once more explicit frameworks are mechanistically fleshed out. However, we do not limit this future development to originating only from robots with physical hardware, perhaps with the goal of mimicking a "real body". These surface-level similarities might eventually turn out to be less critical

to the scientific progress of grounded cognition overall. Some of this exciting and innovative engineering-oriented research will be outlined in the next chapter.

In conclusion, we believe that the focus of the traditional AI on movement and action planning has led to researchers overlooking semantic modelling. We also argue that there is a significant gap in the cognitive science extant literature on grounding cognition mechanistically in real-world visual information. The aim is to sidestep the dilemma of *modeller-defined* and *hand-crafted* semantic input features which are particularly problematic for evaluating cognitive semantic models given the model outputs are reflections of the statistical regularities encoded in the input data. Additionally, manual feature engineering does not facilitate scalable evaluations of cognitive semantic models. Grounding semantic models creates new avenues for assessments using large-scale language corpora and neuroimaging data. Our research objective is further aided by the unabated rise in unstructured data like naturalistic photographs, freely available on the internet, and the increased availability of powerful deep learning algorithms for exploiting the statistical regularities in real-world pictures. Our current thesis brings us back full-circle, from the early pioneering efforts of Sowa's (1976) conceptual graphs to the rise of behaviour-based robotics and subsequent spread of grounded theories in psychology to finally our original research programme of visually grounding semantics using recent advances in artificial intelligence.

Chapter 3

Feature-based and Grounded Semantic Representations

3.1 Abstract

Fodor (1998) claimed that relations between concepts in a semantic system are insufficient for mapping concept correspondences. Goldstone and Rogosky's (2002) study based on the ABSURDIST graph-matching algorithm showed that relations between concepts are critical for representing semantics. Our conceptual replication of Goldstone and Rogosky's finding is based on real-world visual grounding, instead of an arbitrary set of *absolute* (extrinsic) and *relative* (intrinsic) coordinates. We use the *feature-based* approach from Rogers and McClelland (2004) for encoding *intrinsic* representations while using our novel 2D silhouette-based technique for encoding *grounded/extrinsic* representations. Hybrid representations consist of half of the most informative features from both the feature-based and grounded inputs. The same neural network architecture is implemented across all three conditions. Our results support

Goldstone and Rogosky, by revealing that hybrid representations show a markedly slower rate of decline in concept classification accuracy as a function of increasing levels of noise perturbations applied to the hidden layer representations. Grounded representations perform the poorest, while feature-based inputs are moderately tolerant to increasing levels of noise. Hybrid representations are mutually reinforcing and superior to either grounded or feature-based representations in isolation, suggesting a more pluralistic cognitive semantic perspective.

3.2 Introduction

In semiotics, the study of meaning-making, the *semiotic triangle* is a well-known descriptive model of the relationships between real-world *objects* (e.g. a real dog), *symbols* (e.g. the word DOG) and *mental concepts*, first proposed by Ogden and Richards (1923). Semantic representations are conceptual structures of human knowledge and are typically considered to have both symbolic and semantically rich properties (Shallice & Cooper, 2011). In cognitive science, these representations are typically instantiated as interconnected networks of associations. Feature-based representations are symbolic in the sense that their *format* is arbitrarily associated with the real-world objects they represent, per Newell and Simon's (1976) *symbol manipulation perspective*. In contrast, grounded conceptual representations are, at least partially, based on *sensorimotor* (e.g. vision, audition and haptic) and *somatic* (e.g. emotion) states.

A long tradition in computer science and machine learning involves applying noise to feedforward neural networks and observing the subsequent impact on the networks' behaviours. In most cases, the aim of injecting noise into the network is to either improve model convergence or avoid overfitting and thereby increase the generalisability of a given network beyond the training data. This is irrespective of whether the noise is additive vs multiplicative, cumulative vs non-cumulative or even

normally vs uniformly distributed (Jim, Giles, & Horne, 1996). From a psychological perspective, noise tolerance of *artificial neural networks* (ANNs) is seen to correspond to theoretically important aspects of modelling human cognition. For example, in computational neuropsychology (e.g. Hinton & Shallice, 1991), the lesioning of ANNs has been associated with “graceful degradation” of the network’s performance, as opposed to catastrophic effects.

McClelland and Rogers’s (2003) Nature review paper outlined the significance of understanding the dynamics of noise on ANNs that model aspects of semantic knowledge. McClelland and Rogers observed the impact of adding random noise (of varying degrees) to the activations of particular properties of different concepts. The noise was fed into a feedforward neural network trained on labelled data for associating concepts with features and characteristics (e.g. Canary-CAN-Grow, Canary-CAN-Move, Canary-CAN-Fly and Canary-CAN-Sing). Interestingly, they found that the impact of adding increasing quantities of random noise to the inputs of the network led to perturbations in the semantic representation of the network, approximating the impact of destroying neurons associated with representing concepts in the brain. They also discovered that shared properties across concepts (e.g. Canary-CAN-Grow) are significantly more resistant to noise injections than more idiosyncratic properties (e.g. Canary-CAN-Sing), at equivalent levels of noise perturbations. Furthermore, with increasing levels of noise perturbations, concepts become more generalised. For example, concepts typically not associated with their superordinate category’s prototypical properties (for example, Penguin-CAN-Fly) become more activated.

One direct relationship between such computational noise perturbation experimentation using artificial neural networks and cognitive science is the case of *semantic dementia*, which is a progressive neurodegenerative condition related to a deterioration of semantic memory as a result of a loss of neurons associated with semantic memories located

in the anterior and lateral regions on the temporal lobes (McClelland & Rogers, 2003). Increased levels of noise perturbations in ANNs can be equated to the severity of the neuronal loss or brain atrophy in semantic dementia. Interestingly, McClelland and Rogers' (2003) *Parallel Distributed Processing* (PDP) noise modelling could simulate the symptom of patients with semantic dementia, who display a tendency of overgeneralising common words for more specific ones (e.g. "animal" instead of "a rabbit"). In other words, there are real parallels between injecting noise in an artificial neural network and cognitive processes being distorted as a result of the neuronal loss. However, to what extent are the semantic models of human memory as explicated by McClelland and Rogers related to theories of grounded cognition? In the next sections, we will briefly cover a variety of different perspectives that have been particularly influential in accounting for how meaning is processed in the cognitive system.

AI was very much divided into two distinct camps from its inception - *symbolic AI* versus *sub-symbolic* or *neural networks* (Franklin, 2014). Both of these areas were marked with one commonality: they made extensive use of hand-coded and feature-based knowledge representations, in the sense that mental representations of concepts were captured using arbitrary and discrete notations. We provide an overview of disembodied symbolic models and sub-symbolic models, as well as grounded cognitive developmental robotics. This is followed by a presentation of some original computational research that investigates the robustness of *grounded, feature-based* and *hybrid* representations.

3.3 Disembodied Symbolic Models

In *artificial intelligence*, abstract and amodal symbolic models were the norm well before the term was coined by John McCarty for the 1956 Dartmouth Workshop, featuring Marvin Minsky, Claude Shannon and numerous other pioneers of the field. Other such formalisms were

developed in the mid 30s and early 40s such as Alan Turing's *Turing Machine*, Alonzo Church's *Lambda Calculus* and Emil Post's *Production System* (Franklin, 2014). *Good Old-Fashioned AI*, more commonly known by its acronym GOFAI, a term typically reserved for traditional symbolic AI, is primarily based on a series of programmed instructions operating on a set of formal amodal symbolic representations. The core assumption operating across most GOFAI algorithms is that the storage format of knowledge is inherently meaningless, and atomic symbols are arbitrary and can be manipulated based on a set of formal rules essentially a *grammar*. Although, a gross oversimplification, much of GOFAI can be described as comprised of CONDITION-ACTION or IF-THEN rules that make use of atomic symbols called *productions*. Hence the systems built with this approach are called *production systems*.

In the early cognitive sciences, Newell and Simon's (1972) work underpinned the state-of-the-art thinking of such techniques. They developed dynamic goal hierarchies, without the need for overly explicit goal transitions, through the chaining of goals and sub-goals. This led to other mechanisms being developed to allow artificial cognitive systems to cope with increased complexity. For example, *conflict resolution* mechanisms started becoming increasingly important with the advent of implicit goal hierarchies, which could lead to productions being activated simultaneously as a result of a particular condition (Johnson-Laird & Shafir, 1993). Newell and Simon (1980) were the first to characterise and formalise the notion of the *Physical Symbol Systems* (PSS), that had clear parallels to Turing machines, but that also significantly extended the original ideas of symbolic manipulation by grounding it in the cognitive sciences. Notions such as *tokens* - arbitrary symbols, *token-relations* (for example, based on proximity) - and expressions for token handling were deployed to account for the production of symbolic interactions, in other words, *meaning*. They also proposed that such computations within a PSS would be sufficient for intelligent behaviour to emerge. Related concepts such as production rules,

probabilistic mechanisms, frames, fuzzy sets, and more complex data structures were gradually incorporated in GOFAI. A constant across all these developments was the core reliance on knowledge being represented using logical formalisms (e.g. *predicate calculus*), whereby atomic symbols in and of themselves were only meaningful in their interrelations with other equally meaningless atomic symbols.

One core assumption of this disembodied symbolic approach is that the physical structure of the brain is not essential when constraining the types of computational mechanisms responsible for human cognition. At this stage, it would be useful to explore some of the overarching details of a Turing machine, as this is mostly seen as synonymous with computation, which in turn is seen to be a general theory of cognition, more commonly referred to as the *Computational Theory of Mind* (Ludwig & Schneider, 2008). This has implications not only on our discussion and review of symbolic and sub-symbolic models but also grounded versus feature-based cognitive models.

Alan Turing formally demonstrated that deterministic formal systems could be represented computationally with a so-called *Turing machine*, a mathematical processor defining an abstract machine consisting of tape of infinite length (subdivided into cells) and a head able to read and write symbols onto this tape. The Turing machine is always in one of an infinite set of possible states, and its next state is entirely dependent on its current state and the symbol on the tape currently being read. Furthermore, which new symbol to overwrite the current state's symbol, which cell to move to next and which new internal state to enter, are all determined based on a *table of rules* (Haugeland, 1985). Concerning cognitive science, this implies that the physical structure of a particular machine (organic or inorganic), is irrelevant for furthering our understanding of cognition because cognition is computation (Ellis & Humphreys, 1999). This is in stark contrast to grounded views of cognition, where the hypotheses are

mostly based on the specific physical embodiment of the intelligent agent and their situational grounding.

Nevertheless, numerous critics object to the general idea of Turing's thesis and Simon and Newell's PSS hypothesis, whereby symbol manipulation is the critical characteristic of intelligent systems. For example, Harnad (1990) argued that cognition, and in particular, semantic cognition from the perspective of physical symbol system theorists, suffers from the consequence of circular definitions akin to trying to find the "true meaning" of a particular word in a dictionary only to "bounce" from one entry to another, which is the symbolic equivalent of a *merry-go-round*. At this stage, we will omit a review of much of GOFAI, even cognitive architectures such as ACT-R, SOAR or COGENT, as this would exceed the current scope of highlighting computational models of semantic cognition and more pertinently, would not further our discussion of grounded versus feature-based knowledge representations. Instead, we will turn our attention towards historical and current formalisms and models of semantic cognition.

The *Arbor porphyriana* (tree of Porphyry, *see figure 3.1*) marked the birth of semantic networks as a form of knowledge representation in the third century A.D., by the Greek Philosopher Porphyry (234 - 305 A.D.), who extended Aristotle's syllogistic reasoning with the explicit demarcation of relationships using tree branches and leaves (Sowa, 1979). In particular, Porphyry's tree consisted of taxonomies starting from the most generic level (SUBSTANCE → material → BODY vs SUBSTANCE → immaterial → SPIRIT), to intermediate levels (e.g. BODY → animate → LIVING vs BODY → inanimate → MINERAL) and the final level (ANIMAL → rational → HUMAN vs ANIMAL → irrational → BEAST). Although Porphyry's tree was successful in clearly defining relations, one of its disadvantages was the lack of formal notations. Charles Sanders Peirce (1839 - 1914), one of the founders of semiotics, the study of sign systems, introduced *existential graphs* as a means for exhaustively outlining first-

order logic. This graphical representation could also be translated into predicate calculus, hence having the advantage of formalising a vast array of everyday problems mechanistically using a finite series of truth-functional steps, leading to a binary final-state (Bundgaard & Østergaard, 2007).

Ross Quillian's (1967) work on semantic associative networks using spreading activations marked the first foray into representing concepts in parallel across various nodes, based on taxonomic hierarchies between concepts being represented by nodes, and their associations being captured by the types of relations explicitly encoded in the network links. Collins and Quillian (1969) extended this notion of hierarchies being the fundamental structure of cognitive models of semantics by providing mechanistic explanations of then well-known behavioural studies exploring reaction time differences between verifying sentences such as "a terrier is a dog" and "a terrier is a mammal" using the ISA relation. They proposed that the faster reaction time for the *dog* condition versus the *mammal* condition is a result of having to traverse one (terrier ISA dog) versus two ISA nodes (terrier ISA dog ISA mammal) in the conceptual network.



Figure 3.1: Arbor porphyriana (tree of Porphyry). Source: Hellström (2012).

Although these historical and more recent examples of structuring meaning have been widely influential, a key shortcoming of these methods is the lack of focus on how the human cognitive system acquires these representations. The main theoretical emphasis has been historically placed on the representational structure as opposed to knowledge acquisition and induction. A new general theory to address this shortcoming stems from the development of *Latent Semantic Analysis*.

3.4 Latent Semantic Analysis

The quintessential and widely popularised approach to modelling semantics from a distributional perspective is *Latent Semantic Analysis* (LSA), outlined in a seminal publication by Landauer and Dumais' (1997). LSA statistically realises the *distributional hypothesis*, that meaningfully related words tend to appear in similar contexts (Wittgenstein, 1953). LSA is a theory and practical algorithm for inductive knowledge acquisition and representation. At its most fundamental level, LSA is both a theory and a statistical technique for meaning representation relying on a large *bag of words* (BoW), typically consisting of large volumes of unstructured text data. In mathematics, a *bag* is termed a *multiset*, where a set is allowed to have duplicates and the order is irrelevant. For example, if $\text{bag } A = \{\text{dog}, \text{dog}, \text{cat}, \text{rabbit}\}$ and $\text{bag } B = \{\text{cat}, \text{dog}, \text{rabbit}, \text{dog}\}$, A and B are equivalent. LSA measures the frequency of these words occurring across many documents. LSA is a corpus-based approach to modelling semantics and is technically not a cognitive model per se.

Latent Semantic Analysis is a type of *vector space model* (VSM), where documents are represented as points in a vector space, with points closer together being more semantically similar, while points further apart are more dissimilar. The core theoreticians behind the development of LSA (e.g. Landauer & Dumais 1997; Landauer, 2002) even describe their approach as a new paradigm to cognitive science itself. The focus is not on

collecting data under strict laboratory conditions, as is typical for much behavioural research, but under naturally realistic conditions, given that the underlying data fed into LSA is generated naturally from people doing tasks in the real world.

One of the most significant advantages of VSMs over GOFAI systems of semantic modelling is the ability to generate knowledge representations without the need for tedious hand-coded atomic units of knowledge and their interrelations by manually creating rules and numerous ad-hoc adjustments to handle exceptions (Norvig, 1992). VSMs were initially created for applications such as information retrieval and many of their core principles led to the creation of search engine algorithms such as Google's *PageRank* (Manning, Raghavan, & Schütze, 2008). It is likely that VSMs will only become increasingly important as online search gradually transitions from keyword searches to semantic web search.

There are numerous misconceptions of what LSA is comprised of, both regarding a theory of meaning as well as its relatively sophisticated algorithmic implementations, with many psychologists simplifying it to a mere proximity matrix (Landauer et al., 2007). This is an unfair treatment of the elegance and simplicity underlying LSA and also misrepresents one of the core aspects of the technique, namely its ability to make inferences about in-depth relations and acquire knowledge through *inductive inference*, hence the emphasis on *latent* meanings. Landauer and colleagues have stressed on numerous occasions that they do not assume the mathematical details of LSA to be mechanistically related to cognitive or neural processes.

LSA typically generates thousands of semantic dimensions when capturing the statistical variabilities in word co-occurrences. These dimensions correspond to the underlying semantic features embedded in the text, known as the *latent semantic dimensions*. One has to trade off between having too few (< 300) and too many dimension (> 2,000). Having an insufficient number of dimensions leads to a poor representation of the semantic content of the documents in the raw corpus, while too many

dimensions will lead to over-fitting and idiosyncratic associations emerging, that is more a result of the specific texts being used as opposed to the underlying semantics within and between passages (Landauer, Foltz, & Laham, 1998).

Using Marr's tri-level of analysis (Marr, 1982), starting with (i) *computational* (*what* does a system do), (ii) *algorithmic* (*why* does it do these things), and (iii) *implementation* (*how* does it do these things), the present author argues that LSA occupies Marr's first level of abstraction - the computational layer. However, others (e.g. Landauer et al., 2007) have proposed that LSA may provide a narrow perspective at the algorithmic level. Indeed, one of the undeniable successes of LSA is its reasonably long history of successfully modelling a range of behavioural findings from the cognitive sciences, ranging from lexical decision tasks to semantic priming studies (Landauer et al., 1998), and its use in symbolic AI systems and contemporary machine learning.

When LSA models are used for comparisons with behavioural data, the higher-dimensional vectors are used to represent different words, and words are compared to each other using the *cosine* so that the range is between -1 and 1 (inclusive). For illustrative purposes, the closest words associated with the target word *cup*, and their corresponding cosine could be as follows: *mug* (0.84), *glass* (0.54) and *bowl* (0.24), which indicates that *mug* is semantically more similar to *cup*, followed by *glass*, and *bowl*. Interestingly, one of the surprisingly simple ways of extending such comparisons beyond single words, for example, to sentences, is to sum the vectors across all the words in a sentence and then compare the cosines of the summed vectors representing the sentences.

A core point of comparison between LSA and human knowledge representation is based on word-word similarity ratings. Based on Anglin's (1970) empirical research on word sorting using concept and grammatical relations, it has been shown that children and LSA have a word-word similarity correlation of 0.50, though this decreases to 0.32 for adults,

because parts-of-speech structures become increasingly important, and are beyond the mechanistic scope of LSA (Landauer et al., 1998). One intriguing test of LSA in a human domain was conducted by Landauer et al. (1998) by training an LSA model on corpus data consisting of a first-year undergraduate psychology textbook (Myers, 1995), and then evaluating the model based on four-option multiple-choice questions used for two cohorts of undergraduate psychology students. The LSA model scored 60% in this semantic domain test, which was significantly above chance, and was sufficient for passing the tests, although slightly below the average score of the undergraduate students. This proof-of-concept experiment indicates that by adding up the individual word-level vectors, more complex semantic chunks (e.g. questions) can also be represented using LSA.

More recently, Jones, Kintsch and Mewhort (2006) have shown how LSA's cosine similarity measures are predictive of *lexical priming effects*, a "bread-and-butter" technique used in psycholinguistics and cognitive science, where recognition of a word is faster and more accurate when preceded by a related word (for example, *doctor-nurse*). Nicodemus et al. (2014) even used LSA in a clinical neuroscientific study investigating the genetic pathways related to *category fluency* - a commonly used neuropsychological task. In this task, participants are required to name as many instances of a category (e.g. *ANIMAL*) in a short period, and the key measures are the overall number of correct words listed (e.g. *DOG*) and the number of incorrect words listed (e.g. *CAR*). However, Nicodemus and colleagues also used LSA to generate clusters of words for target categories based on the higher-dimensional meaning space, for representing the cognitive search processes of navigating the semantic space from the target category. Based on the behavioural data from the category fluency task and the LSA-derived category fluency measures, Nicodemus et al. narrowed the search for common genetic variations in schizophrenia, called *single nucleotide polymorphisms* (SNP), across 665 participants and discovered three candidate SNPs for further research.

In providing this highly selective glimpse into the extant literature of Latent Semantic Analysis, we have seen that LSA has not only become one of the most influential theories and models of semantics but is also being used across a wide range of clinical and commercial applications, such as genetic profiling and information-retrieval systems.

3.5 Disembodied Sub-Symbolic Models

Sub-symbolic models of semantic processing emerged as a direct challenger to more traditional symbolic models, many of which required time-consuming and cumbersome explicit hand-coding of symbolic facts in the form of rules (and exceptions) such as propositional networks. As a general rule of thumb, sub-symbolic models tend to be based on representations that have an element of statistical learning and as a result are also more likely to be associated with continuous values, unlike symbolic models, which commonly have discrete values representing specific informational states. The present section will focus on *artificial neural networks*, also known as *connectionist networks*. Other disembodied sub-symbolic models like *genetic algorithms* and *fuzzy classifiers* will not be covered as they are not widely used for modelling cognitive semantics.

Traditional computers, but also the majority of GOFAI and symbolic models of cognition, rely on serial processing, where a single computation occurs at any given moment, and linear chains of computations unfold. Having said this, it is worth noting that some production systems do indeed have weaker elements of parallel or cascaded processing such as Soar's *elaboration phase* and ACT-R's *spreading activation*. However, like the human brain, artificial neural networks operate in parallel, in other words, processing can co-occur across the system, which is why these types of models are also widely known as *parallel distributed processing* (PDP) models. The fundamental principle with connectionist networks is that a large number of elementary processing

units can lead to highly complex and intelligent emergent behaviours akin to biological systems relying on billions of neurons (Goodfellow et al., 2016).

Neural networks and related approaches like *deep learning* have been immensely successful and popular in AI. Deep learning typically consists of many hidden layers and more complex learning rules and procedures involving representations generated with Boltzmann machines. Recent developments over the last decade have seen deep learning systems significantly outperform once cutting-edge AI technologies on core tasks like image and speech recognition (LeCun et al., 2015). These approaches can solve highly complex problem solving on board games like *Go* (Silver et al., 2016), which was once thought to be the pinnacle of all AI challenges. Some estimate that deep learning alone could be worth \$17 trillion US dollars². For purposes of brevity and relevance, we will omit technical details such as *AlphaGo*, Google DeepMind's Go-playing AI, being a hybrid system based on deep neural networks and tree search. However, much of this hype can be traced back to developments in the underlying mathematics of these deep networks and the availability of cheaper and more powerful processors, in particular, the *Graphical Processing Units* (GPUs) and *Tensor Processing Units* (TPUs) being utilised. Although these developments are impressive, they are less important for evaluating historical and current developments of PDP models in cognitive science, even though this is likely to change in the future. We will now explore some of the underlying principles associated with connectionist models developed in cognitive science, with a particular focus on semantic representations.

Connectionist networks are usually composed of a set of input and output nodes, with a set of hidden nodes (or several layers of hidden nodes) sandwiched in-between. These nodes are usually connected, and in many

² Source: ARK Investment Management LLC, Global Federation of Exchanges.

cases, fully inter-connected between the layers, which all have weights representing the strength of associations between the nodes. Each of these nodes has a threshold for firing (activating the next node), and a particular type of activation function that computes and transforms the net inputs coming into that particular node. A common activation function widely used in cognitive science modelling is the *sigmoid activation function*, an S-shaped function, where, as the net input to that node increases, the probability of that node firing also increases. Neural networks are typically initialised with normally-distributed random weights, although, this is a widely-debated and very active area of mathematical and statistical research, see Yam and Chow (2000) for a discussion. A common approach to learning in simple 3-layered feed-forward networks is the back-propagation of error algorithm, which takes place through several *epochs*, each following a six-step process. First, stimuli activate the input nodes. Second, these input nodes activate hidden layer nodes. Third, the hidden nodes activate output nodes. Fourth, outputs are compared to target values and fifth, an error signal is propagated back to input nodes. Lastly, the error signal is used to change the connection weights given a specific learning rate, where a larger rate equals faster learning, with a cost of missing the optimal minima (Goodfellow et al., 2016).

The most influential computational theory of semantic processing was developed within the PDP research programme, the “hub-and-spoke” model (Patterson, Nestor, & Rogers, 2007). This model integrates earlier conceptual ideas postulated by Donald Hebb in the 1940s (Hebb, 1949) with Geoffrey Hinton’s multi-layered connectionist modelling of storing proposition knowledge (Hinton, 1981). One of the first PDP models to be used by the PDP Research Group, which also acts as a foundation for many related models (see Rogers & McClelland, 2004 for a historical overview), as well as the original modelling research presented at the end of this chapter, consists of a simplified neural network model of semantic memory (Rumelhart, 1990). This feed-forward neural network model learns

associations through back-propagation, by associating a set of arbitrary/symbolic stimuli and concept labels, where similar inputs to the model are represented by corresponding hidden layer representations. This PDP computational model demonstrates that traditional taxonomic hierarchies (e.g. Quillian, 1967) can also be represented using sub-symbolic feedforward neural networks.

However, in theory, the PDP framework does not exclude the possibility of grounded representations being used as the inputs. Rogers and McClelland (2004) claim that perceptual similarities can indeed be used as inputs. However, PDP models are currently still based on inputs that are arbitrary and symbolic (Barsalou, 1999). Although, one of the first pioneers to use connectionist modelling to ground categories in perceptual stimuli was Harnad (1992), who used feed-forward neural networks with back-propagation, for category-learning of lines of eight different lengths into three categories of *short*, *middle* and *long*. Based on a set of elegant and straightforward simulations on one-dimensional stimuli, lines being represented as input vectors (though a variety of different formats were used), Harnad (1992) mechanistically extended the original theoretical ideas proposed in Harnad (1990), on the symbol-grounding problem, and offered a tractable toy-model solution to address the challenges of connecting physical objects to symbolic cognitive models. More contemporary research, leveraging this approach and significantly extending it originates from the more recent field of *Cognitive Developmental Robotics*, which will be our final focus in this review of mechanistic models of human semantic cognition. However, we first explore some shortcomings of the connectionist approach.

Despite the success of PDP models in cognitive science, there are three common shortcomings associated with such models. Firstly, from an information theoretical perspective, one might argue that the present distinctions between symbolic and sub-symbolic computational processes are incorrect at worst and superficial at best. This is because, quite

paradoxically, sub-symbolic models like artificial neural networks can indeed be formally expressed by a universal Turing machine, and therefore by definition, are symbolic. Earlier discussions of connectionism being *non-symbolic* (Fodor & Pylyshyn, 1981), have been mainly replaced by the sub-symbolic interpretations initially proposed by Smolensky (1988). Furthermore, like in the case of LSA's usage and interpretation within psychology, neural networks are also not interpreted from a symbolic perspective and are typically described as sub-symbolic in introductory textbooks (e.g. Eysenck & Keane, 2005). The main reason for their sub-symbolic interpretation in psychology stems from neural networks' learning and storing distributed representations, as opposed to having specific tokens representing entire concepts.

Secondly, in spite of broad structural similarities between *artificial neural networks* and *neuronal assemblies* in the brain (e.g. *nodes/neurons* or *weights/synapses*), the differences outweigh the similarities. McClelland, Rumelhart and Hinton (1988) state, despite surface-level similarities and neuro-plausibility of these PDP models, the main appeal is based on providing a mechanistic account of psychological phenomena for furthering our understanding of potential candidate computational structures and processes. Although many researchers use neural networks as a mechanistic proxy to real neural architectures for simulating specific neural circuitry (e.g. Sejnowski, 1981), this is uncommon in cognitive science, where ANNs are more widely implemented as a *behaviour-based* approach to problem-solving. This, however, also leads to the third shortcoming. Interpreting the behavioural characteristics of neural networks can be somewhat challenging in the absence of explicit representations that can be formally analysed. Current cutting-edge hybrid methodologies between ANNs and GOFAI approaches (e.g. fuzzy rule-based systems) are helping to gradually overcome this limitation.

3.6 Grounded Developmental Robotics

In this final overview section, our focus will be on robotic systems inspired by grounded and developmental principles, with a particular focus on the grounding of language in sensorimotor interactions of robots. Developmental robotics is an interdisciplinary paradigm for robotics research and is tightly coupled with practical insights from developmental psychology (Cangelosi & Schlesinger, 2015). First, however, we will aim to provide a highly selective overview of how the field of developmental robotics emerged, in order to illuminate its unique inter-disciplinary origins and its central difference from traditional robotics and AI research.

Robots are synthetic organisms that function autonomously and adaptively under ever-changing environmental demands. The word ‘robot’ originates from the Czech writer Karel Čapek’s play R.U.R (*Rossum’s Universal Robots*), and means “menial labour”. Historically, the origins of this field can be traced all the way back to mechanical mathematics, intricate mechanical ducks (e.g. Jacques de Vaucanson’s eating, drinking and digesting duck) and modern-day factory robots.

Asada and colleagues (2001) claim that developmental robotics is a new paradigm for humanoid robotics because of its focus on sensorimotor contingencies that enables learning to occur in naturalistic settings. However, for a more psychologically relevant perspective on developmental robotics, with a detailed and up-to-date introduction and analysis of the field, Cangelosi and Schlesinger (2015) provide a wide-ranging analysis of developmental psychological phenomena and their role in developing learning mechanisms in artificial systems. Next, we will cover some of the work from Angelo Cangelosi and collaborators in more detail, starting from software-only artificial neural networks to cognitive systems embedded in software simulations and physical instantiations (via the humanoid robot *iCub*).

Cangelosi, Greco and Harnad (2000) pioneered the use of neural networks for hybrid symbolic and sub-symbolic systems to not only ground actual objects, but to use these basic representations to further ground more abstract representations associated with these objects, through a process of *grounding transfer*. In their computational experiments, they used a set of four shapes (circle, ellipse, square, rectangle) defined using a retinal representation coding (Jacobs & Kosslyn, 1994) instead of abstract representations such as circle = {1,0,0,0}, ellipse = {0,1,0,0}, square = {0,0,1,0}, and rectangle = {0,0,0,1} which would be common practice. This more complex representation of the inputs was based on retinal units receiving inputs from partially overlapping receptive fields originating from a 50×50 pixel matrix, where the centre of the receptive fields had a greater influence on the retinal units because of a Gaussian distribution function centred on the receptive field. In addition to this retinal input, there were six additional linguistic and symbolic input nodes, using a one-hot coding scheme, for each of the four shapes and the symbolic designations *symmetric* and *asymmetric*. There were five hidden nodes, and the output nodes mirrored the input nodes. Abstract categorisation (symmetric versus asymmetric) is feasible directly based on the inputs alone, which they term *sensorimotor toil*. Their method of *grounding transfer* used a Boolean combination of grounded category names. This method allowed for the indirect grounding of the abstract concept of symmetry. This study demonstrated that the so-called sensorimotor toil dilemma, a time-consuming and unrealistic process of supervised learning of all concepts, could be avoided via *symbolic theft* because some concepts can be grounded indirectly in raw sensory activations, while others could then be based on a scaffolding built on the grounded categories.

Cangelosi and Riga (2006) extend the grounding transfer mechanism using two robots (one as the demonstrator and the other as the imitator) in a virtual environment, where each robot has 12 degrees of freedom, a humanoid-shaped torso, with two ‘arms’ forming a gripper, on

a four-wheeled robotic chassis. In this study, the back-propagation of error in the neural network of the imitator robot not only learns to correct its motor response (based on joint angles) but simultaneously learns the linguistic labels of the corresponding actions. Therefore, the imitator robot (after training) mirrors the actions of the demonstrator robot and can also perform actions based on their corresponding names. Using a similar *modular neural network* (MNN) approach to Greco, Riga and Cangelosi (2003), the imitator robot's higher-order abstract categories are grounded in the combinations of specific names. Cangelosi and Riga implement a chained approach to sensorimotor grounding, with three distinct stages. In the most fundamental level of grounding, known as *basic grounding* (BG), the imitator robot's eight primary action names (CLOSE_LEFT_ARM, CLOSE_RIGHT_ARM, OPEN_LEFT_ARM, OPEN_RIGHT_ARM, LIFT_LEFT_ARM, LIFT_RIGHT_ARM, MOVE_FORWARD, and MOVE_BACKWARD) are directly based on the robot's actions in the virtual environment. A novel approach to the higher-order grounding of concepts is introduced in this study through the demonstrator robot also providing names of combined actions. The second type of grounding is known as first-order *higher grounding* (HG1), where two basic actions that have been previously trained via BG are combined into a composite behaviour. The example provided by Greco et al. is "GRAB [is] CLOSE_LEFT_ARM [and] CLOSE_RIGHT_ARM" (p. 681). Whereas, the second-order *higher grounding* (HG2) results from the combination of a basic level action (BG) and a first-order higher-order routine (HG1). This research demonstrates not only the feasibility of grounding new actions in direct sensorimotor interactions, although this is still a highly active area of research in AI and cognitive science (e.g. Kiela, 2017; Kiela & Clark, 2017), but exhibits the feasibility of language from others (the demonstrator robot) being leveraged for indirect grounding of actions.

Yamashita and Tani's (2008) cognitive developmental robotics research with a different humanoid robot provides yet another extension of

the original work on grounding transfer mechanisms developed by Cangelosi and Riga (2006). Yamashita and Tani's functional hierarchies of handling objects emerged through the system's self-organisation of temporal sequences of behaviours segmented into discrete and modular units. This is an intriguing prospect for AI research, although it seems that imitation might be a more psychologically plausible mechanism for cognitive modelling purposes (Iacoboni, 2009; Cangelosi, 2010).

Stramandinoli, Marocco and Cangelosi (2017) extend their original grounding transfer mechanism with a more explicit focus on linguistic grounding of abstract words which cannot be grounded in perceptual modalities, to explore the nature of abstract action words through the analysis of the model's internal activation units using Principal component analysis (PCA). To the present author's best knowledge, Stramandinoli et al.'s research was the first set of directional results demonstrating that grounding abstract action words can be achieved by reusing (at least partially) the sensorimotor representations, in support of a grounded perspective on higher-order cognition. This indicative finding is based on a qualitative analysis of a 3-dimensional PCA plot of the hidden activations of the various types of network training, and the significant overlap in the hidden unit activation values when comparing sensorimotor trained representations with abstract action verbs (e.g. *MAKE*).

An interdisciplinary perspective comprised of psychology, neuroscience, and linguistics coupled with the cognitive modelling and the burgeoning domain of cognitive developmental robotics is critical for a deeper understanding of the interplay between action and language, and also to "disentangle ambiguous issues, provide better and clearer definitions, and formulate clearer predictions" (Cangelosi & Borghi, 2014, p. 346). Cangelosi and Borghi concretely further the debate on grounded versus abstract representations concerning action and language by identifying three phenomena. First, natural language in conjunction with sensorimotor and emotional information play a role in conceptual

representations. Second, the *mirror neuron system* (MNS) is related to action and language processing. Third, developmental robotics, using artificial embodied agents - both physical robots or software emulations - can play a role in identifying the mechanistic interplay in the integration between action and language. Cangelosi and Borghi (2014) also highlight that cognitive scientists and philosophers who are influential in developing grounded cognition as a discipline (e.g. Clark, 2008; Barsalou, 2008), are reconsidering their view on cognition being purely grounded in sensorimotor activity, and are also considering the role of language. This marks a shift in the direction of the grounded cognition research paradigm because it could help overcome the unproductive impasse on embodied versus disembodied cognition and help generate new empirical and computational research directions.

In a related research endeavour on grounding mechanism, De La Cruz et al. (2014) explored the feasibility of embodied models via epigenetic and grounded approaches to modelling elementary mathematical cognition, in particular, handwritten digit recognition based on the widely used MNIST machine learning benchmark dataset. This MNIST training data consists of 28-by-28 pixel images split across ten classes (0 - 9). The embodied modelling was run using the virtual simulator of the iCub humanoid robot and three modules: an auto-encoder neural network for processing visual information (*visual module*), a feed-forward neural network for controlling iCub's fingers (*motor module*), and a generalised regression model for associating the different classes of numbers to particular finger configurations (*visuo-motor association module*). These modules were pre-trained independently, before their merging, after which the entire network was trained using backpropagation. Although the iCub platform's hands each have 9 degrees of freedom (DoF), only seven were used in this study. The thumb, index and middle fingers each have 2 DoF, while the pinky and ring fingers, which are glued together, have only 1 DoF. De La Cruz et al. (2014) compared the digit recognition rate between

the embodied version of the model and a disembodied version without finger configurations in the form of proprioceptive sensory data and found that the embodied model performed significantly better (greater classification accuracy), with a difference of 7.4% between the embodied and standard model.

In the final section of this chapter, we will experimentally explore grounded, feature-based and hybrid representations. Despite a plethora of computational research on different types of representations, as can be seen from the overview of abstract, both symbolic and sub-symbolic, and embodied developmental robotics, there has been limited focus on comparing these different representations. In other words, models have been created for two divergent approaches, with a lack of empirical and modelling work at the intersection of these polarised theoretical perspectives. Therefore, we will explore the robustness of these distinctive representational formats in a proof-of-concept computational experiment focusing on noise tolerance.

3.7 Computational Study I: Noise Tolerance

3.7.1 Theoretical Background

Human beings have a remarkable ability to categorise objects in the real world based on both perceptual features (e.g. a dog has legs, or a bird has feathers) and relations between concepts (e.g. bread and butter), which is typically known as the *conceptual web account* (Quine & Ullian, 1970). Goldstone and Rogosky (2002) investigated connecting concepts to either each other, an external source, or both, using a *constraint satisfaction network* for translating between two artificial conceptual systems and a graph matching algorithm called ABSURDIST (Aligning Between Systems Using Relations Derived Inside Systems Themselves). In their constraint satisfaction network, each node represents a given hypothesis regarding a

correspondence between systems A and B, while the interconnections between the nodes constitute the constraints. Therefore, the ABSURDIST network does not learn any associations per se but is merely used to store a set of hypotheses regarding the similarity of the two inputs of coordinates representing the different conceptual systems. Externally-grounded representations consisted of absolute coordinates, while internally-grounded ones were based on the relative distances between the coordinates themselves. Principally, the ABSURDIST algorithm is not a computational model of cognition but an abstract information theoretical construct for investigating how conceptual systems can have meaningful correspondences. They extend Fodor's (1998) notion of conceptual identity between different conceptual systems being a prerequisite for successful mapping by proposing similarity itself would be sufficient. In other words, the concept of a *teacup* does not have to be identical between two people in order for their respective conceptual systems to contain the concept, as long as the concepts each play an equivalent role within their respective systems (Goldstone & Rogosky, 2002).

The ABSURDIST algorithm evaluates systems based on comparing the Euclidean distances of two matrices and finding correspondences between the two. It is evaluated based on its *tolerance to distortion*, by gradually perturbing the coordinates of the "concepts", with noise sampled from a Gaussian distribution with a mean of zero. Goldstone et al. tested the ability of ABSURDIST to find correspondences between two idealised and simplified conceptual systems as a function of ever-increasing amounts of noise being injected into the network and tested how internally-, externally-grounded or a combination of the two impacts ABSURDIST's conceptual mapping performance.

The ABSURDIST simulation results showed that two conceptual systems do not need to contain identically defined concepts in order for these to be successfully mapped to each other. This argues against Fodor's (1998) claim of concept identity being critical for conceptual

correspondences. Goldstone et al. further discovered that for conceptual translations, internally-grounded concepts (based on relative Euclidean distances between coordinates) performed better than externally-grounded concepts (absolute coordinates), but that a combination of both internally- and externally-grounded concepts had the best performance. Hybrid stimuli showed the most gradual decline in accuracy between the systems as a function of increasing noise). These findings suggest an interesting hypothesis, namely that, there might be a mutually reinforcing relationship between intrinsic and extrinsic information. This would have strong implications for radically embodied accounts and grounded-only or distributed-only accounts of human cognition. The results suggest that a combination of extrinsic and intrinsic relations for representing concepts is advantageous.

In the next section, we extend the ideas of Goldstone and Rogosky, by modelling artificial conceptual systems using feedforward neural networks from a cognitive modelling perspective as opposed to an ABSURDIST-like general graph matcher. Consequently, the neural networks in our simulations below will learn conceptual representations using training data consisting of feature-based, grounded and hybrid representations. Based on Palminteri, Wyart and Koechlin's (2017) delineation of computational modelling approaches in cognitive science, Goldstone and colleagues' study can be seen as an analytical approach, whereas our adaptation represents an instantiation of cognitive hypotheses. This not only allows us to model human cognition and thereby draw cognitively-relevant mechanistic implications but also to bridge the mostly independent computational modelling work spanning the areas of traditionally feature-based semantic modelling, and the more recent and extensive cognitive developmental robotics research, inspired by theories of embodied and grounded cognition.

3.7.2 Methodology

3.7.2.1 Feature-based Data

In this study, 20 amodal representations originate from a direct adaptation of Rogers and McClelland's (2004) extended training corpus of animal and plant features (see *figure 3.2*). Inputs in this feature-based representation capture the conceptual similarity based on the types of features selected in the original model developed by Rogers and McClelland (2004).

	Pine	.	.	Salmon
ISA... Living_thing	1	.	.	1
ISA... Plant	1	.	.	0
ISA... Animal	0	.	.	1
ISA... Tree	1	.	.	0
ISA... Flower	0	.	.	0
Is... Pretty	0	.	.	0
Is... Big	1	.	.	0
Is... Living	1	.	.	1
Is... Green	1	.	.	0
Can... Grow	1	.	.	1
Can... Move	0	.	.	1
Can... Swim	0	.	.	1
Has... Skin	0	.	.	1
Has... Roots	1	.	.	0

Figure 3.2: Example format of the feature-based representations from Rogers and McClelland (2004).

3.7.2.2 Grounded Data

Although the PDP approach implemented in Rogers and McClelland (2004) could be used to represent perceptually-inspired semantic representations by replacing the original context-property conceptual pairs with perceptually-inspired characteristics, this would not be based on any current theories or hypotheses of grounded cognition. However, one of the most influential grounded hypotheses is Barsalou's *Perceptual Symbols System*, which is prominently associated with silhouettes of concrete objects like *CHAIR*, as that is the example used by Barsalou (1999). In developmental psychology, there is well-established empirical evidence (Landau, Smith, & Jones, 1988, 1992; Imai, Gentner, & Uchida, 1994) that lexical categorisation is learnt from distinguishing between

shapes based on the saliency of object functions and parts. More recently, a large team of *Google DeepMind* researchers (Ritter et al., 2017), used state-of-the-art one-shot learning models trained on the ImageNet dataset, to show that their computational model also had a *shape bias* akin to the findings from the developmental psychology literature, indicating that shape, as opposed to colour, is the dominant factor for categorising objects. As a result of these empirical and computational experiments, the grounded data format used in this study is *object form*, more specifically, the silhouette extracted from naturalistic images of the 20 different concepts being represented. This is a gross simplification of genuinely grounded perceptual inputs but is used in the present proof-of-concept study to ensure that grounded and feature-based representations are well-matched in their informational content.

The grounded input is captured by first extracting the outlines of the 20 objects from standard JPEG colour images. In theory, state-of-the-art *Convolutional Neural Networks* (CNN) would be the optimal class of algorithms for extracting the most semantically-sensitive information from these images, especially when combined with other current trends in machine learning such as *pooling* and *averaging* across pixels from this type of image data (Goodfellow et al., 2016). However, that would change the architecture of the network and would therefore not be a controlled comparison with the feature-based data described previously. Therefore, the 2D outline of the image is transformed from a 2D to a 1D format by finding the medial point of the image (x_c , y_c), and then iteratively (i) calculating the Euclidean distance (d_i) between this point and successive contour points in a counter-clockwise pattern (see *figure 3.3*). This transformation has three fundamental shortcomings, which are: (i) unequal vector lengths across the 20 concepts (depending on the number of pixels in the silhouette of the image), (ii) exceeding the number of data points feeding into the 50 input nodes of the network, and (iii) having data intervals that do not fall within the 0 and 1 range as in the feature-based

model. These shortcomings are overcome by applying a smoothing function to the raw 1D vector and then sampling 50 equally-spaced intervals to output a 1D vector of length 50, which we then rescale to make the maximum Euclidean distance equal 1.

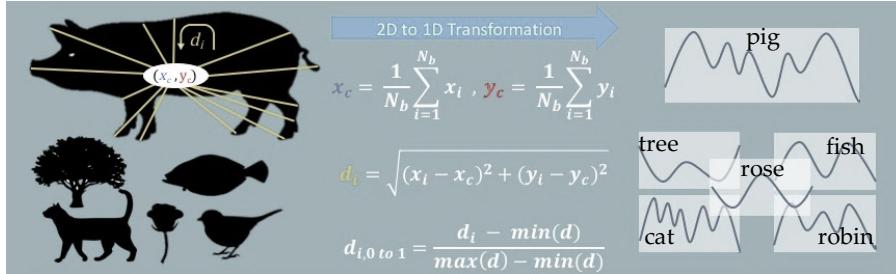


Figure 3.3: Silhouettes and their corresponding silhouette profiles based on the 2D to 1D transformation.

3.7.2.3 Simulation Details

We use an artificial neural network consisting of two hidden layers (60 nodes in layer 1 and 30 nodes in layer 2) using the back-propagation supervised learning algorithm. The network weights for all neurons are initialised randomly using a Gaussian distribution, and the back-propagation algorithm is run for 10,000 iterations, using a logistic activation function, with a learning rate (α) of 0.01 and momentum (β) equalling 0.9. This ensures that the final network weights (post-training) are the result of gradual learning processes, where the initial random values are small and have a weak influence on the target values. A small learning rate is traded off with a larger momentum parameter to ensure that the global minima is not missed but simultaneously speeding up the rate of convergence (Goodfellow et al., 2016). However, given that both parameters typically range between 0 and 1, the modelling is skewed towards a slower algorithm with more stable learning. Finally, the network error is measured using the *Sum of Square Error* (SSE). We follow the approach laid out in Rogers and McClelland (2004) and ensure that during network training, each epoch consists of presenting each stimulus once in a randomised order since the weights are updated after every stimulus. The core assumption is that in most ANN models, minor adjustments are made to the network, simulating

everyday experiences, which gradually build towards a distributed knowledge representation.

The same neural network architecture is implemented for comparing feature-based, grounded and hybrid representations' tolerance to noise perturbations. Suitable data formats are selected for each condition, while not biasing the number of input representations. Therefore, in addition to the same neural network architectural and learning parameters, each model also consists of the same number of data inputs and outputs (20 target concepts). Furthermore, it would be a biased comparison of the hybrid representation against the two others if the hybrid inputs consisted of twice as much information or were trained on an artificial neural network with different architectural parameters (e.g. 2×50 inputs). In the case of the feature-based and grounded data representations, we use a constant number of network inputs ($n = 50$). In the case of the hybrid data, the methodological challenge of representing both feature-based and grounded data formats as network inputs without doubling the total number of input nodes (from 50 to 100) is overcome by sub-sampling 50% of the most informative features (least correlated) within the feature-based and grounded stimuli. Therefore, the total number of hybrid inputs remain the same during network training across all three conditions.

Model	Architecture	No.	Learning	Momentum	Error
		Iterations	Rate (α)	(β)	
Feature-based	L1.60 : L2.30	10,000	0.01	0.9	0.00929
Grounded	L1.60 : L2.30	10,000	0.01	0.9	0.00943
Hybrid	L1.60 : L2.30	10,000	0.01	0.9	0.00914

Table 3.1: Model parameters and sum square error for ANN training of feature-based, grounded and hybrid models.

The noise perturbation is applied to the network using a 3×21 factorial design, with three categories of data representations (feature-

based, grounded and hybrid) and 21 levels of noise increment from a factor of 0 (no noise) to 1 (maximal noise), with increments of 0.05 noise. The random noise is sampled from a uniform distribution and is combined with the input data to simulate the loss of neurons associated with the semantic information. Each network is trained 1,000 times at every noise increment, resulting in a total of 63,000 ($3 \times 21 \times 1000$) neural network models, each being run for 10,000 epochs. This is required to generate 1,000 predictions for every noise increment so that reliable estimates across the predictions can be made to generate an overall accuracy level for a given type of network (based on input representations) at a particular noise increment. In this experiment, in addition to the two core manipulations, firstly the format of semantic representation, and secondly, the level of noise, there is a further peripheral manipulation of investigating how noise impacts semantic representations as a function of the number of concepts being represented. This additional objective aims to better understand the scaling potential of grounded representations, which has not been previously explored, unlike the case of distributed representations, where simulations (Goldstone & Rogosky, 2002) have shown that networks become more efficient and more tolerant to noise as they scale up. Finally, another aim of this study is to explore the resultant semantic representation following the training of the neural network with grounded representations, and to evaluate the grounded network's ability to classify stimuli that are not part of the original 20 concepts used during training.

3.7.3 Results

We explore the noise tolerance of the neural networks trained on feature-based, grounded or hybrid representations. Classification accuracy (%) is measured based on the neural networks correctly classifying the 20 concepts according to the diagonal of the *confusion matrix* between observed and predicted classes. Unsurprisingly, figure 3.4 shows that as the level of noise perturbation increases from a factor of 0 to 1, in 0.05 increments, for

all types of data formats, there is a gradual decline in accuracy, though not reaching chance-level benchmarks across the three models (for 20 concepts, at chance-level, the benchmark is $100/20 = 5\%$).

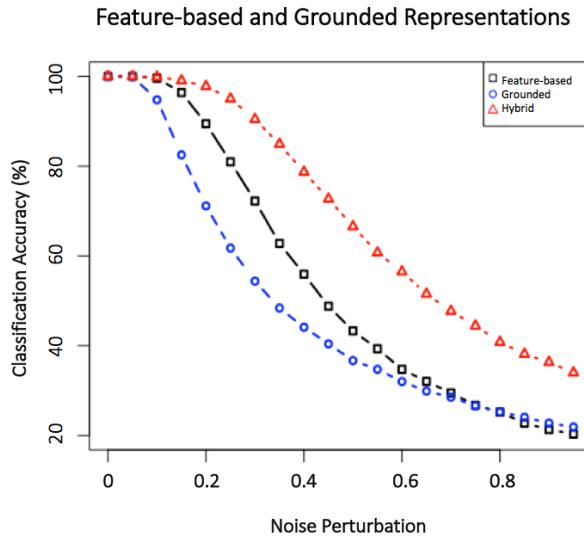


Figure 3.4: Performance of the neural networks trained on feature-based, grounded and hybrid representations as a function of noise. Error bars are not included as the error margins are too small.

Interestingly, however, when looking at the moderate range of noise (between 0.2 to 0.6), feature-based representations are more tolerant to noise than grounded representations, although with a higher level of noise (between 0.65 and 1) this difference disappears and both representations are equally impacted by noise perturbations. The most interesting finding of these simulations is the superiority of hybrid over both grounded and feature-based representations once noise perturbation reaches 0.15, which is maintained up to a noise factor of 1, where grounded and feature-based models level-off at 21% while the hybrid model plateaus at 38%, an almost two-fold superiority in noise tolerance.

Previous computational experiments (Goldstone & Rogosky, 2002) have shown that as intrinsic systems scale, their tolerance to noise increases almost exponentially. In the current simulation, we first explore whether there is also an advantage to noise tolerance as the number of concepts increase in grounded models. Second, we explore the nature of this increase using an experimental design of 7×21 factorial combinations, with seven

categories of number of concepts (in our simulations: 2, 3, 4, 6, 10, 15 and 20 concepts) and the same 21 levels of noise increments used previously. Once more, 1,000 neural network simulations with differently seeded initialisation weights are run for each one of these combinations ($7 \times 21 \times 1,000 = 147,000$ models) before estimating the average classification accuracy across the 1,000 models for each experimental condition.

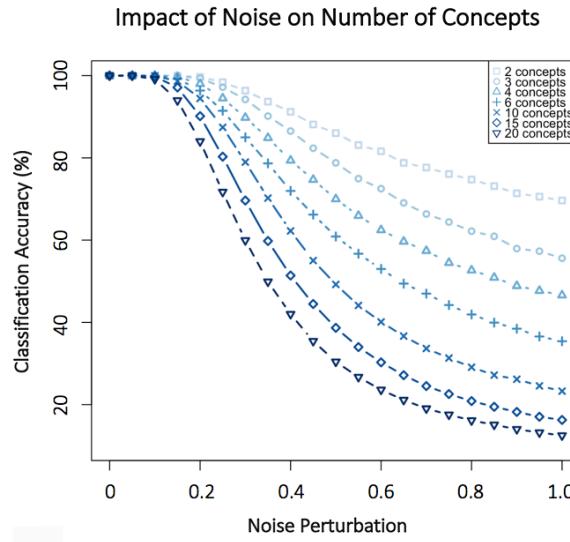


Figure 3.5: Classification accuracies of the grounded neural network.

In *figure 3.5*, we can see that the grounded network's classification accuracy increases as a function of the number of concepts being represented (compared to relative chance-level benchmarks). This is likely to be due to the interrelations between the different inputs being more distributed for a larger number of concepts. Even at a noise perturbation factor of 1, there is a positive relationship between the number of concepts represented in a network and level of classification accuracy compared to chance-level benchmarks. For example, with 4 concepts, the model is 1.76 times better than chance (chance-level: 25% vs model: 44%), while with 15 concepts this increases to 2.55 times (chance-level: 6.67% vs model: 17%), and with 20 concepts this increases to 3 times as high as by chance alone (chance-level: 5% vs model: 15%).

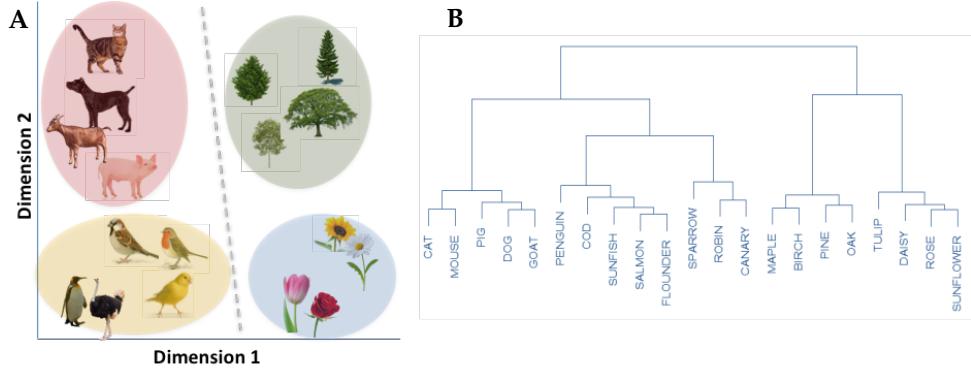


Figure 3.6: A Multidimensional Scaling (MDS) representation of the grounded semantic space (A), and a hierarchical cluster plot of the same grounded semantic space (B). Please note that for the MDS plot, the fish category has been excluded for depicting the other associations more clearly.

Lastly, from *figure 3.6* we can see that our silhouette-based grounded representation successfully captures the taxonomic properties of all 20 concepts based exclusively on perceptual inputs. The superordinate categories *animal* and *plant* are split first, followed by further divisions between *mammals* and *birds*, and *trees* and *flowers*, respectively. Moreover, the grounded neural network can also successfully generalise by meaningfully representing the novel concept *OSTRICH*.

3.8 Discussion

Our proof-of-concept simulations of noise perturbation on grounded, feature-based and hybrid representations presented in this chapter highlight the potential benefits of pluralistic semantic representations. In conjunction with the extant literature on cognitive developmental robotics, our study suggests that it may be helpful to move away from the black-or-white debates on grounded versus amodal cognitive representations. This debate on “are concepts grounded?” might be more helpfully reformulated as “to what extent are concepts grounded?”. Additionally, the present work supports Lenat et al.’s (1991) prediction that as the so-called conceptual web increases in size, it becomes

easier for semantics to emerge from rich intrasystem relations, as larger semantic networks are more tolerant to noise.

A new finding in this study is that semantic features can be learnt purely through the bottom-up grounding of perceptual characteristics, even when exclusively using shape information. Without including higher-order semantic characteristics in the training data (e.g. ISA...Tree|Plant|Animal or CAN...Fly|Sing|Move|Grow), the present grounded feedforward neural network model shows how generalised features in the raw perceptual inputs themselves can provide sufficient statistical regularities for meaningful semantic classifications at a conceptual level. A critical difference between the conceptual relations codified in Rogers and McClelland (2004) and the present study is that we avoid the *semantic circularity* in modelling higher-order relations (e.g. plants vs animals) by excluding these explicit distinctions in the grounded training data as this only includes a 2D-to-1D transformed vector-representation of the silhouette, and no other meta-data. However, one might argue that Rogers and McClelland (2004) only used their example as a proof-of-concept illustration for outlining the general principles underlying the emergence of semantic relations. We agree with this interpretation but would add that current approaches to connectionist modelling in cognitive science predominantly make use of hand-coded features in the training set, while the present example uses information naturally available in our environment. This allows for the relatively straightforward and efficient scaling of the present mechanistic approach for representing semantics grounded exclusively in shape information. Furthermore, using the same neural network trained on the grounded data, classifying novel concepts that were not part of the original data set is demonstrated by the new concept *ostrich*. This novel concept is not only correctly “placed” in the *bird* category, but also near the *penguin* (see *figure 3.6*), another atypical representation of a bird (based only on shape). The *Multidimensional Scaling* (MDS) plot in *figure 3.6* reveals the broad categories

that emerge from the neural network trained on grounded data, while the *hierarchical cluster analysis* shows how the divergence of concepts is taxonomically aligned to their superordinate plant and animal category despite no explicit information being made available during the training phase of the network.

Bridges between perceptual and conceptual systems are not new (see Goldstone & Barsalou, 1998), although the present noise perturbation experiments show that in parallel distributed systems, a sensory reductionist perspective where semantics is entirely contingent upon raw sensory data is unlikely to lead to the stability required for conceptual systems like semantics and language to develop. Therefore, it is highly unlikely that grounded and feature-based representations are mutually exclusive but are more likely to be *mutually reinforcing*, based on their distinctive contributions to the semantic system.

Despite a long history of symbolic and grounded accounts of semantic cognition being represented as contrasting philosophical and empirical positions, there might be mutual benefits for both types of representations, given that their unique statistical regularities provide a more robust and intertwined semantic representation. This view is also supported by detailed reviews of empirical studies of semantics. For example, Dove (2011) defends this pluralistic account of cognition based on the ability for such representations to have inferential advantages for both linguistic and sensorimotor semantic codes. Dove argues that natural language can leverage the grounded perceptual codes to extend our cognitive reach via symbolic mechanisms such as *deductive* and *inductive reasoning*. Similar views, albeit from a more computational perspective, are provided by Clark (2008, p.47), where he emphasises the fruitful synthesis of context-free, amodal and arbitrary symbols being able to manipulate non-arbitrary, modality-rich and context-sensitive grounded representations.

3.9 Summary

In this chapter, we started with a brief overview of a distinct set of abstract symbolic and sub-symbolic models and grounded models from a cognitive developmental robotics perspective. Although these categories are well-known in cognitive psychology, these distinctions are not entirely accurate when taking into account the technical details of the models themselves. However, we have outlined the models' distinctions in this manner to reflect the dichotomies present in contemporary cognitive science. We have also seen that computational models tend to either be agnostic to claims of grounding or be inspired by embodied/sensorimotor theories, as is the case with developmental cognitive robotics. However, there is a lack of computational modelling research looking at the role of scene-based grounding of semantic representations.

In our proof-of-concept simulations of grounded, feature-based and hybrid representations we demonstrate for the first time, that the statistical regularities present in the silhouette of an object are sufficient for creating semantic embedding spaces well-documented in the cognitive science literature based on hand-coded features. The grounded model is sufficient for not only classifying trained concepts but also meaningfully classifying new concepts that are not originally part of the training set.

Finally, from the simulations, there is supporting evidence in favour of hybrid grounded and feature-based inputs being superior to either format in isolation. This is in broad support of a pluralistic model of conceptual processing that transcends grounded versus feature-based debates and seeks to explore the interplay of both types of information.

Chapter 4

Extending Symbol Interdependency: Perceptual Scene Vectors

4.1 Abstract

Louwerse (2011) empirically advances the symbol interdependency hypothesis by demonstrating the importance of statistical regularities in linguistic surface structures. Symbol interdependency posits that meaning extraction attributed to embodied representations or algorithms should instead be attributed to language. In a series of 7 experiments we find evidence for language surface structures encoding meaning best when sufficiently constrained by modeller-determined feature sets, with performance deteriorating for randomly selected language surface structures. Furthermore, Latent Semantic Analysis' meaning encoding improves as weaker dimensions are removed. These findings collectively indicate that although language is important, increasing the relevance of linguistic statistical regularities is also critical. Our novel approach, *Perceptual Scene Vectors* (PSVs) use object co-occurrences from images to

automatically extract strong associative and taxonomic relationships more successfully, measured both qualitatively and quantitatively, with an original application of a cluster-correspondence metric. PSVs encode meaning without modellers hand-coding relevant features, which provides an ecologically valid approach to extending symbol interdependency beyond language and partially solving the *relevance problem* in semantics by grounding meaning extraction in real-world visual scenes.

4.2 Introduction

Human symbolic cognition is commonly argued to be mediated and moderated by language itself, essentially proposing that language shapes our thinking. This general idea is typically summarised as *linguistic relativity* or the *Sapir-Whorf* hypothesis (Gipper, Sapir, & Whorf, 1972), while the stronger view is called *linguistic determinism*, which claims conceptual processing is entirely dependent on the language people are exposed to and habitually use. The linguist Benjamin Lee Whorf, proposed that language shapes conceptual categories, which in turn shapes the way we conceptualise our world (Whorf, 1956). In linguistic determinism, the relationship between language and cognition is tightly coupled while the association between cognition and the world is weaker. In contrast, in cognitive science, the perception is that thought is more tightly coupled with the world, while language augments cognition, like a technology, by extending non-linguistic conceptual representations (Holmes & Wolff, 2010).

The philosophical origins of linguistic relativity can be traced back to the two founders of *Semiotics*, Ferdinand de Saussure (1857 - 1913) and Charles Sanders Peirce (1839 - 1914). De Saussure (1916) distinguishes between *sound-images/signifiers*, linked to the real-world and symbolic entities that are carriers of meaning and the *concept/signified* which represents the mental concept itself. These two theoretical primitives, in

conjunction, yield a *sign*, referring to symbolic interdependency being a core aspect of human meaning-making. However, the more prominent theory within Semiotics is Pierce's *meaning triad* (see figure 4.1), which consists of the following three components: (i) *object*, (ii) *interpretant* and (iii) *representamen*. The *object* component has the most straightforward interpretation in the sense that it is directly grounded in a physical object, even though it should not be confused with the real object, because according to Pierce, it is a mere approximation of the true object in the real world. Therefore, it is also commonly referred to as the *semiotic object*. The *representamen* is the amodal representation such as the word DOG or {01100100 01101111 01100111}, the binary representation of the word DOG. Lastly, the *interpretant* refers to the "pure conceptual content" or semantic aspect of the symbol. As in De Saussure's theory, Peirce's theory also emphasises symbol interdependency, although the mechanistic details remain highly underspecified (see Merrell, 1997 for a review).

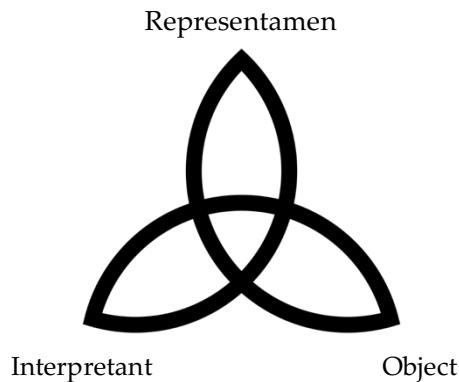


Figure 4.1: Charles Sanders Peirce's notion of Sign.

Related to Pierce's meaning triad, is Paivio's *Dual Coding Theory* (DCT), also cited in Louwerse's (2007) original postulation of symbol interdependency and therefore sufficiently important for a quick review. According to DCT, there are three planes of semantic abstractions, which are (i) *representational*, (ii) *referential* and (iii) *associative*. The representational level approximately corresponds to Peirce's semiotic object level and representamen level because of the encoding of both linguistic (e.g. the

word BOOK) and non-linguistic (e.g. picture of a book) perceptual information. The referential level creates links between verbal (called *logogens*) and non-verbal representations (called *imagens*). Lastly, the associative level consists of an interconnected web of relations exclusively between amodal word-level representations. However, unlike Peirce's semiotic triad, Paivio's DCT goes beyond the merely descriptive-level theorising of meaning systems and suggests a mechanistic theory relating the different levels to existing ideas of the semantic memory system, for example, the referential level being equated to working memory, whilst the associative stage represents long-term semantic memory. However, Louwerse (2011) highlights that Paivio's DCT is distinct from Louwerse's symbol interdependency theory in that the latter proposes that language encodes perceptual information, which in turn mediates both verbal and nonverbal cognitive processes (p. 280).

More recently, cognitive anthropologists have amassed a range of observational findings, purportedly in favour of linguistic relativity, by uncovering thought processes they claim are associated with specific linguistic phenomena. For example, Everett et al. (2005) outlines the linguistic constraints on colour perception based on a small remote group of indigenous people called the *Pirahã*, located in the Amazon rainforest in Brazil. The *Pirahã*, a monolingual hunter-gatherer tribe, find it easier to both remember and recognise particular colours based on the availability of their colour vocabulary. Similarly, Gordon (2004) supports the linguistic relativity hypothesis by demonstrating that the *Pirahã*, who have a very limited vocabulary for numbers are supposedly limited to approximate rather than exact numerical competencies. Gordon claims that this supports the *Sapir-Whorf* hypothesis, since it appears to be the case that speaking a language without number words has a direct causal influence on the way speakers of *Pirahã* perceive exact quantities.

However, experimental evidence from Frank et al. (2008) provides some of the strongest evidence to date contradicting linguistic relativity,

because in a range of studies they find the Pirahã tribe successfully perform tasks requiring conceptual processing of exact numerical magnitudes. Frank and colleagues argue that given the well-known limitations of visual and auditory short-term memory (Baddeley, 1988), the accuracy required for the numerical magnitude tasks of Gordon (2004) significantly exceeded working memory capacities and more realistically required symbolic encoding of the numbers themselves, in the absence of a numerical vocabulary for larger quantitates.

Frank et al. ran a set of cardinality matching tasks, wherein the addition or subtraction of one quantity leads to a difference in the sets to be matched against the control set. The Pirahã are relatively successful at this despite not having the linguistic machinery for representing cardinality and/or addition and subtraction, which Frank and colleagues interpret as the Pirahã having a mental representation of *one* despite not having a word for this number concept. Therefore, in direct contradiction to arguments put forward by Gordon (2004) from the perspective of linguistic relativity, the concept of exact cardinality is not dependent on language. However, having number words can be a cognitive technology that facilitates the ability to remember and operate on exact numbers. These results do not support a strong Whorfian perspective. In fact, this demonstrates that language is an abstraction, or in the words of Frank et al., "a cognitive technology" (p. 2), that builds a scaffolding for more complex abstract mental operations, but does not mediate all mental activities.

Other evidence (e.g. Gilbert et al., 2006) has shown that in the perceptual domain of colour, language can significantly facilitate performance in discrimination tasks through the use of linguistic memory aids and preferred mental processing routes using words as placeholders for concept-environment pairings. Thus, language seems to merely provide alternative, and at times, more efficient processing routes for the encoding and decoding of perceptual experiences as opposed to fundamentally changing the phenomenological nature of cognition itself.

4.3 Evidence for Symbol Interdependency

Louwerse's (2007) *symbol interdependency hypothesis*, states that linguistic processing can be both embodied and disembodied based on their respective reliance on modal and amodal semantic processing. The philosophical debate on whether language encodes our perceptual information has a long tradition dating back to Gorgias of Leontini (485 - 380 BCE), who went so far as to argue that the physical world can only truly be phenomenologically experienced and understood through language. However, according to more recent philosophers like Immanuel Kant (1724 - 1804), the structure and function of language is just one of several dimensions that contributes to the human experience of the world (Jarratt, 1998). Thus, language is seen to have a moderating as opposed to mediating influence on the conceptualisation of reality through the primary senses.

One of the simplest and consistent dichotomies found in both language and action is the presence of nouns and verbs, respectively corresponding to objects and actions. Monaghan, Chater and Christiansen (2005) conducted a series of corpus analyses for discriminating between grammatical categories of words (e.g. nouns versus verbs) based on 16 phonological (e.g. phoneme length, syllable length, presence of stress) and distributional cues (co-occurrence with words such as ARE, NO, YOU, THAT's and ON). Based on the discriminant analysis run, Monaghan et al. report that 67% of nouns and 71% of verbs are correctly classified using only distributional and phonological cues. In other words, based on phonological and distributional linguistic information alone, it is possible to distinguish between verbs representing actions and nouns representing objects. However, Monaghan and colleagues do brush aside the significant proportion of the variability unexplained by phonological and distributional information alone, given the lack of other semantic dimensions they could have tested to explore the relative importance levels.

Further evidence directly supporting symbol interdependency's assumption of language encoding perceptual information, arises from a recent study conducted by Walker and Parameswaran (2019), which elaborates on Walker's (2016) notion of *sound symbolism* - sound affords expectations about the salient aspects of particular words. Walker and Parameswaran conducted a counterbalanced between-participants experiment where participants put their hands separately into two thick denim bags and grasp onto the two cylindrical objects of different weight. The lighter cylinder weighed 44g and the heavier one 190g. The experimenter then informed the participants that the objects were *kipli* and *moma*, without revealing which object corresponds to which name. They were then subsequently asked to decide what each of the objects were called. Lastly, this pre-test condition was followed by two additional conditions to isolate a contribution from vowel and consonant contrast cases. As predicted, in the *pre-test* condition Walker and Parameswaran found that in 80% of the cases, the heavier object was named *moma*, whereas, *kipli* was judged to be the name for the lighter object. Also, as predicted, in the *vowel contrast test*, in 81% of the cases, the names *kipli* and *mimi* were assigned to the lighter object, whilst *kopa* and *moma* were attributed to the heavier object. Collectively, these results suggest that heavier objects are associated with vowels that symbolise "less pointiness" (e.g. *kopa* and *moma*) and have open vowels, such that the tongue is placed further away from the roof of the mouth (Eide & Gish, 1996). In contrast, vowels with "more pointiness" (e.g. *kipli* and *mimi*) are judged to be associated with lighter objects. Walker and Parameswaran propose that this supports language encoding cross-sensory correspondences in the form of sound symbolism based on open vowels representing heaviness by reference to visual roundedness. Therefore, like Monaghan et al., Walker and Parameswaran also support language and the associated linguistic machinery as subsuming sensorimotor affordances - a core tenet of symbol interdependency.

Theoretical evidence referenced by Louwerse (2011) in support for symbol interdependency stems from Christiansen and Chater's (2008) novel theoretical account for the following question "why is language so well suited to being learned by the brain?" (p. 490). Although, Christiansen and Chater's account is heavily critiqued in the subsequent open commentary by numerous cognitive scientists, their perspective is sufficiently novel and relevant to symbol interdependency to warrant a brief overview. Christiansen and Chater argue that language is akin to an organism and by extension faces the same evolutionary pressures as other organisms do. Intriguingly, according to their argument, language has evolved to survive in the environment of the human brain under selection pressures faced by human learning. Therefore, in contradiction to grounded cognition theories, they claim that language is not actually an arbitrary abstraction or meta-cognitive function but is shaped by perceptual, learning, processing and pragmatic constraints. Language encodes the structure of thinking and is not a mere adaptation for communicative purposes.

Symbol interdependency is theoretically appealing due to the mutually reinforcing effect of hybrid representations and the evolutionary metaphor outlined by Christiansen and Chater (2008). Goldstone and Rogosky (2002) demonstrate the theoretical system-level advantages of intrinsic and extrinsic representations, which, in chapter 3 of this thesis, we also find to be the case for feature-based and grounded cognitive representations. Furthermore, symbol interdependency is also empirically appealing based on findings such as discriminating between actions and objects using linguistic structures (Monaghan et al., 2005) and the effects of sound symbolism investigated by Walker and Parameswaran (2019).

4.4 Surface Semantic Analysis

Louwerse (2011) states that the importance of statistical regularities in linguistic structures is underestimated in both symbolic and grounded theories of cognition and that the formation of semantic representations, typically attributed to either specific statistical processes, or grounded representations should be explained by language itself. The emphasis is particularly on mining the semantic representations found in the first-order or so-called surface structure of language. First-order co-occurrences are the simplest form of frequentist descriptive statistics on words co-occurring together. We term Louwerse's approach to extracting meaning from text, *Surface Semantic Analysis* (SSA), with the aim of positioning it beside the better-known *Latent Semantic Analysis* (LSA). This new approach has a strong ancillary claim of embodied representations being mapped onto language. The approach is based on previous research (e.g. Louwerse & Jeuniaux, 2010) showing a dissociation between deep and shallow language processing. Deep processing referring to low-ambiguity linguistic situations, whereas shallow processing is both underspecified and incomplete. The default mode for language users is shallow processing. In contrast, deep processing is a more deliberate and effortful mode of linguistic processing, akin to the widely popularised *system 1* (fast "autopilot") versus *system 2* (slow and logical) dichotomy of human decision making (Kahneman, 2011). Louwerse's (2011) theorising is highly relevant to the present thesis, and its importance to the debate on grounded and amodal representations is succinctly summarised in the following excerpt:

"The prediction that language encodes perceptual information has important implications for the symbolic and embodied cognition accounts presented earlier. For the symbolic cognition account, it means that results obtained from LSA can also be obtained through non-latent patterns in the language surface structure. For the embodied cognition account, it means that results attributed to perceptual simulations can be traced back to language itself." (p.7)

According to Louwerse (2011), distributional linguistic information is sufficient for accounting for the statistical variabilities found in sensorimotor data. Louwerse suggests that even though some researchers (e.g. Landauer et al., 1998) focus on particular mathematical and computational details of algorithms for explaining meaning induction (Latent Semantic Analysis), symbol interdependency rests exclusively on language. Louwerse makes a strong assertion for meaning being directly extracted from language itself, and that the current focus on particular types of algorithms such as LSA or even neural network models is misguided. This contradicts Landauer and Dumais' (1997) emphasis on a powerful mathematical analysis implemented in LSA that leads to meaning induction.

Louwerse (2011) asks whether or not first-order occurrences will result in identical results to that of latent patterns. If LSA and SSA both yield the same results, then mere surface-level measures are sufficiently powerful for extracting meaning from language, and provide a simple "lower-bound" measure for what humans can extract from language (p. 13). Conversely, Louwerse reasons, if LSA produces different results to SSA, then first-order co-occurrences are not sufficiently information-rich for meaning extraction. In his first computational experiment Louwerse (2011) used the classic Rogers & McClelland (2004) verbal descriptors (names of birds, fish, flowers and trees) along with 26 features (e.g. bark, branches, ... white, and yellow) to create a 16×26 matrix as the input to multidimensional scaling (MDS). The original matrix is generated using LSA cosine values trained on the Touchstone Applied Science Associates (TASA) corpus. These values are subsequently transformed into Euclidean distances. The MDS coordinates from the x- and y-dimensions are plotted for qualitatively exploring the semantic space (*see figure 4.2a*). In a similar fashion, another 16×26 matrix is generated, but using the frequency counts in the *Web 1T 5-gram corpus*, created by Google, which contains English word n-grams (unigrams to five-grams) and their observed

frequencies. The matrix is an input for MDS analysis in order to visualise the verbal descriptors (*see figure 4.2b*). Intriguingly, this was the first study showing that semantic results obtained by LSA can also be reliably replicated using mere first-order n-gram co-occurrences. Based on these results, Louwerse proposes the need to reconsider the results of embodied cognition studies using linguistic stimuli from a disembodied perspective, given that language can encode embodied relations and perceptual information.

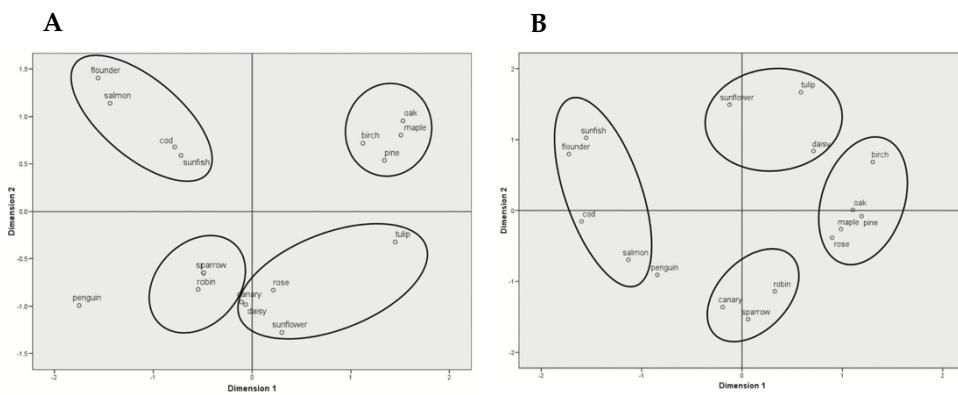


Figure 4.2: MDS plots from Louwerse (2011) of the 16 verbal descriptors \times 26 features used in Rogers and McClelland (2004). (A) MDS plot based on LSA analysis (B) MDS plot based on first-order co-occurrences.

4.5 Computational Experiment: Perceptual Scene Vectors

4.5.1 Theoretical Background

The common themes of the reviewed empirical and computational studies focusing on the interplay between language and cognition are as follows: (i) language contains sufficient sensorimotor encodings for conceptualising reality, (ii) disembodied and embodied cognition are mutually reinforcing, and (iii) there has been too much focus on algorithmic or embodied details of cognition and not enough on language. In this section, we propose to extend the original symbol interdependency hypothesis beyond language, to include other modalities of concept

representation, with a particular focus on our daily stream of contextual information mined from the environment. Much of this information for the normally-sighted population is visual, which has inspired our approach.

Our focus on ecologically valid stimuli based on a realistic scenario is three-fold. Firstly, even though Louwerse (2011) has demonstrated that both LSA and first-order co-occurrences based on n-grams are suitable for taxonomic and associative meaning extraction, a cognitively plausible account of how these associations are acquired or stored in the first place remains elusive. We argue that the proximities used for the verbal descriptors adapted by Louwerse (2011) are based on a set of highly structured big data representation, which is an implausible semantic memory format, despite being exceptionally well-suited for computational meaning extraction. Secondly, the features used in Louwerse (2011) and cognitive modelling more broadly (e.g. Rogers & McClelland, 2004) are hand-coded in the sense that a human modeller manually selects the relevant features to include. Louwerse (2011) implements both the LSA and the first-order co-occurrence models this way in order to replicate a well-known and widely used toy semantic model. However, a general criticism aimed at the wider cognitive modelling community is that reliance on such *a priori* features inevitably introduces the confound that the model reflects the modeller's intuitions of a cognitive phenomenon. This brings us to the third and most important reason for wanting to use ecologically valid inputs for cognitive modelling. Most models of semantic cognition are unable to acquire the relevant associations in the first place. In the case of traditional symbolic models (e.g. Cooper & Shallice, 2000), the triggering functions for particular schemas need to be hand-coded *a priori*, whilst in the case of language-based models (e.g. Louwerse, 2011), the 16×26 matrix needs 26 features to be defined manually. In both cases, one sidesteps one of the key challenges posed by cognitive models of semantics - selecting the most *relevant* semantic features from a collection of hundreds of potential contenders based on task- and context-sensitivity. This is arguably one of

the most critical, yet overlooked, first steps to realistically representing meaning.

There is a need for a computationally explicit model of semantics grounded in acquiring and selecting the most suitable associations from the environment automatically, without human modellers hand-coding a series of features based on time-consuming and non-parsimonious trial-and-error iterations. In order to computationally implement such a model, our original approach for the cognitive scientific study of conceptual processing extracts contextual information from images of naturally occurring scenes, and generates a cognitive data representation, *Perceptual Scene Vectors* (PSVs). PSVs are a type of data representation (a 1D tensor) that capture the statistical regularities of the environment within which an object commonly occurs, so objects that tend to co-occur more often with other objects in a variety of prototypical contexts are more likely to be related to one another. Therefore, contextual coherency in PSVs across multiple exemplars is defined by homogenous distributions and smaller cosine distances between the vectors, whereas incoherent contexts are defined by a set of PSVs with larger cosine distances and heterogeneity in their distributions. The technical implementation details of PSVs is outlined in the modelling section below. However, we will first address the theoretical justification for introducing PSVs and the importance of context more generally and its relevance to conceptual processing.

Why are naturalistic visual contexts important for the study of semantic processing? Based on a review of the extant literature one would have to conclude that ecologically valid stimuli have not been a priority for much of semantic cognition research - both in terms of the empirical literature, based on isolated and decontextualised objects, and computational models, reliant on toy datasets of isolated concrete objects. There are some notable exceptions, as we discussed in chapter 2, when outlining the field of developmental robotics, although traditionally that has been more widely associated with engineering as opposed to cognitive

science. The lack of ecologically valid research on context-based semantics is especially surprising when taking into account the early discovery of *context-dependent memory* - that memory encoding and decoding are subject to environmental factors. Pessin (1932) was probably the first researcher to study the impact of context on both memory and behaviour. In this early study, two conditions are evaluated, one consisting of both visual and auditory stimulation and the other being a quiet environment. In the noisy condition, participants, rather unsurprisingly, spoke more loudly, but also made more overt movements than in the quiet condition. Similarly, in Egstrom et al.'s (1972) study of underwater task performance, they found strong evidence in favour of underwater training being superior to dry-land training for an underwater task. Therefore, the context specificity of memory-encoding during the learning phase is seen as a critical aspect of human learning.

Why has cognitive science research on semantics ignored context? We conjecture that for empirical research, strictly controlled lab conditions are easier to achieve when objects are isolated from their context. While in the case of computational modelling, a relentless focus on the allegedly "key" aspects of a problem are typically a by-product of *reductionism*. In the real world, however, if one were to imagine seeing a slice of toast, it would typically be accompanied by other relevant objects (e.g. a loaf of bread, butter, knife) and contexts (e.g. kitchen). It would be unusual to imagine the slice of bread floating in mid-air void of all other objects in the background. Although, this is precisely the way visual objects are treated in most experiments. Many objects in our surroundings are in fact in the same semantic space and yet physically appear very different (e.g. *laptop* and *computer*), whilst there are numerous instances of the opposite, where they appear to be physically very similar but are in almost orthogonal semantic spaces (e.g. *real handgun* and *toy gun*).

A classic finding within cognitive psychology, is the example of the word BREAD automatically priming for related words like BUTTER (e.g.

Cramer, 1970). *Semantic priming*, is a cognitive phenomenon where a target word (e.g. TRUCK) is processed faster when preceded by a conceptually associated prime (e.g. CAR) as opposed to an unrelated prime (e.g. TIGER). Although, the facilitation effect of word recognition for BUTTER followed by the prime BREAD is a linguistic example, equivalent findings have also been recorded using pictorial or iconic stimuli (see McNamara, 2005, for a detailed review).

In relation to our current discussion of the importance of context on conceptual processing, a more directly relevant phenomenon, is *contextual priming*, which has been widely documented since the 1970s (e.g. Taylor & Juola, 1974; Sanford, Garrod, & Boyle, 1977; Smith, Theodor, & Franklin, 1983). More recently, the phenomenon has also been documented in a range of domains associated with complex human behaviours. For example, contextual priming has been shown to impact evaluations of ambiguous products based on adjacent advertisements (Yi, 1990), legal proceedings by influencing juries and other stakeholders involved in legal proceedings (Fraser & Stevenson, 2014), and even impact whom people voted for depending on where they voted. For example, Berger, Meredith and Wheeler (2008) found that people assigned to a school polling station were more likely to support pro-school funding initiatives than those whose polling station was assigned to a church. Berger et al. found this to be the case even after controlling for demographics, political views and geographical location. These findings collectively demonstrate that contexts should not be overlooked as non-essential characteristics to be excluded from investigations of conceptual processing given its importance to real-world scenarios.

Torralba (2003) demonstrated in a computational model the ability to account for *contextual priming*, where the context of an object provides a superior prime for the identification of the target object, compared to the intrinsic features of the target prime. Torralba argued that real-world scenes are governed by strong configurational properties, that are

particularly useful for identifying objects based on these arrangements. In a related study, Chun (2000) demonstrated that visual contextual information constrains where we look, what to expect and also surprisingly guides our attentional spotlight, which enables us to process visual scenes more effectively. This has been shown to be both learnt and executed implicitly (Chun & Jiang, 1998). In this chapter, we build on the idea of the importance of contexts for object identification and extend it more broadly within the novel context of semantic memory and operationalise the idea through the development of PSVs.

Direct evidence on the importance of visual contexts for concrete objects originates from a phenomenon known as *boundary extension*, where participants not only recall the set of objects they actually saw in a naturally occurring scene, but also extrapolate their recall to other objects that they would expect to see in the natural scene (Intraub et al., 1996). Gottesman and Intraub (1999) have articulated this form of memory distortion as inaccurately recalling *wide-angle views* of close-up scenes. Since the participants in these studies are only provided with the “zoomed in” perspective, their extrapolation to contextually relevant concrete objects is likely to be based on semantic scene associations.

The above studies demonstrate that contextual processing is important for attentional and perceptual processes. However, is it genuinely important for furthering our understanding of higher-order conceptual processing? Bar (2004) reasons that in order for contextual information to be useful to semantic associations, the information extracted from contexts has to be processed quickly enough to enable other cognitive processes to benefit from this rich information. Biederman et al. (1974) conducted four experiments with jumbled and coherent real-world scenes, with participants having to select one label between two options that best describes the scene. Participants were able to extract semantically meaningful information at about 100ms based on selecting scene-relevant labels. Thus, one may even argue that the context is processed before the

object itself is identified. Even though object identification has been historically seen as a serial bottom-up process, there is neuropsychological and functional anatomical evidence of top-down activation of visual processing. Based on the empirical work of Schyns and Oliva (1994), Bar (2004) theorises that this fast processing of contextual information is enabled through global cues being dependent on low spatial frequencies (blurry images showing proportions of “recognisable blobs”), whereas the details of the actual objects themselves are high spatial frequencies analysed during later stages of visual processing (p. 621).

Here, we argue that we see the world as a successive array of scenes, wherein objects (in high spatial frequency) are embedded in a magnificently rich and yet low spatial frequency surroundings with other related objects. We walk into a supermarket and our senses, especially the visual system, are bombarded with a plethora of sensations of hundreds of objects meshed together in a rich tapestry. All this information hits us in a neither serial nor purely parallel manner, but in the form of a continually cascading information stream. In fact, there are well-known theoretical reviews and studies (e.g. Biederman, 1972; Biederman et al., 1982) that have argued that representing and processing concrete objects in naturally occurring groups facilitates the recognition of other concrete objects typically found in the same contexts. A second argument we make here is that objects in our scenes do not inhabit isolated islands; instead, they form a rich and meaningful tapestry of associations, which should be exploited for the study of semantic cognition. Thirdly, on the basis of our first two arguments, we hypothesise that the rich interconnectivity of the objects captured by exploring visual scenes leads to a process of *grounded symbol interdependency*.

Grounded symbol interdependency is a generalisation of Louwerse’s original hypothesis; in our view, an extension to account for non-linguistic concept representations. Finally, we propose that the common associations for both concrete and abstract concepts emerge, to a

large extent, not merely from feature lists or language alone but from task- and context-specific pairings. This not only facilitates perception, but also enables efficient conceptual processing through the increased availability of heuristic shortcuts by a process of *conceptual integration* - the processes by which concepts are linked with one another through PSVs.

In order to integrate concepts, one has to first acquire them. Interestingly, concept formation in children is highly dependent on the ability to visually explore the environment. Comparative studies on congenitally blind and sighted children (e.g. Jaworska-Biskup, 2011) have typically tested school-aged children, using batteries of concepts based on colours, fauna and flora and inanimate objects, and have shown significant limitations in congenitally blind children when it comes to proper concept formation and understanding. The studies also indicate that blind children depend more on contextual cues for successful conceptualisation instead of linguistic structure.

The central hypothesis of the present study is that our surroundings contain sufficient statistical regularities for explaining much of the variability (at least partially) in our semantic space. In algorithmic terms, this amounts to the claim that images of objects in their natural surroundings contain statistical patterns that are sufficiently predictive of the dimensions determining our conceptual space. Therefore, we claim that our cognitively plausible grounding of semantics in visual scenes, will demonstrate computationally the theoretical adequacy for extending symbol interdependency beyond language.

4.5.2 Experimental Hypotheses

Based on the theoretical overview of the importance of context in the study of human conceptual processing, we propose a series of hypotheses, each of which will be evaluated in the present series of computational experiments.

- I. Semantic maps will show **very strong** meaningful associative and taxonomic associations when LSA and 26 n-gram features (hand-coded) are used.
- II. Semantic maps will show **very weak** meaningful associative and taxonomic associations when LSA and 26 random “noun features” are used.
- III. Semantic maps will show **strong** meaningful associative and taxonomic associations when LSA and 26 random “verb features” are used.
- IV. Semantic maps will show **weak** and random associative and taxonomic associations when only LSA 300 dimensions are used.
- V. Semantic maps will show **stronger** and more meaningful associative and taxonomic associations when LSA dimensions are gradually reduced (300 -> 150 -> 50).
- VI. Semantic maps will show **the most meaningful** associative and taxonomic associations when automatically extracted Perceptual Scene Vectors (PSVs) are used.

4.5.3 General Methodology

As in the work of Louwerse (2011), the above hypotheses all relate to results which can be explored qualitatively using MDS plots of various distance matrices. These may also be explored quantitatively through comparisons to an independent ground truth.

4.5.3.1 Verbal descriptors and features

In computational experiments 1-5 of this section, we use the original set of verbal descriptors and features used by Rogers and McClelland (2004). In experiments 6 and 7, we use a new set of verbal descriptors, for which we are not specifying any features as these experiments will respectively be using only LSA dimensions and PSVs (see *table 4.1*).

Categories	Original		New
	Verbal Descriptors	Features	
Animal	PINE	Pretty	tiger
Bird	OAK	Big	elephant
Fish	MAPLE	Green	zebra
Flower	BIRCH	Red	rabbit
Plant	ROSE	Yellow	hamster
Tree	DAISY	White	desk
	TULIP	Twirly	chair
	SUNFLOWER	Grow	drawer
	ROBIN	Move	lamp
	CANARY	Swim	computer
	SPARROW	Fly	butter
	PENGUIN	Walk	jam
	SUNFISH	Sing	toast
	SALMON	Leaves	knife
	FLOUNDER	Roots	plate
	COD	Skin	car
	CAT	Legs	bus
	DOG	Bark	cycle
	MOUSE	Branches	bike
	PIG	Petals	train
		Wings	
		Feathers	
		Scales	
		Gills	
		Fur	
		Living	

Table 4.1: 20 verbal descriptors × 26 features used in Rogers and McClelland (2004); Louwerse (2011) used the first 16 verbal descriptors (excluding mammals). The final column consists of the new verbal descriptors used in experiments 6 and 7.

4.5.3.2 TASA Corpus

All language-based spaces in this study are generated using the TASA (Touchstone Applied Science Associates, Inc.) corpus, which consists of 37,651 individual documents and 92,393 unique terms. TASA is used in the development of the world's most exhaustive educator's word frequency guide, used by publishers for assessing the vocabulary of different textbooks. The corpus itself consists of newspaper articles, novels and online sources. We use a particular corpus library available for the R statistical programming environment (Günther, Dudschig, & Kaup, 2016), which also includes the *Web 1T 5-gram*, consisting of 1 trillion unigram to five-gram tokens used for running LSA. This library is known to have some deviations from the benchmarked LSA spaces of the TASA corpus from the LSA Research Labs in Boulder, USA. Therefore, in order to replicate the current computational models, it is necessary to use the *LSAfun* package in

R. LSA default parameters are used throughout the simulations, unless stated otherwise.

4.5.4 Experiment 1: 20 concepts \times 26 n-gram features

4.5.4.1 Objective and Methods

In our first computational experiment, our objective is to conceptually replicate Louwerse's (2011) first analysis using semantic associations of 16 verbal descriptors \times 26 features, which we extend to 20 verbal descriptors used by Rogers and McClelland (2004). The verbal descriptors are used to extract the relevant n-grams from the 1 trillion records in the *Web 1T 5-gram* dataset. Following the details of Louwerse's experiment, we run classical multidimensional scaling (cMDS) on the matrix and transform the 5-gram log frequencies into a matrix of scaled Euclidean distances. This experiment is aimed at testing hypothesis 1, that *semantic maps and hierarchical clustering will show very strong meaningful associative and taxonomic associations when 26 hand-coded n-gram features are used*. The theoretical reasoning resting on symbol interdependency as outlined in Louwerse (2011).

4.5.4.2 Results

Our first experiment reveals that semantic maps and hierarchical clustering show very strong meaningful associative and taxonomic associations when 26 hand-coded n-gram features are used. In *figure 4.3a* we see the categorisation of trees, flowers, birds, fish and mammals. However, the corplot does show weaknesses in coherence of the flower and mammal categories as the groups, despite being visible, are quite weak. This is also reflected in hierarchical cluster analysis (see *figure 4.3d*), where flowers and mammals form unusual groupings, for example, DAISY is grouped together with DOG and PIG. However, superordinate taxonomic

relations (e.g. PLANT vs ANIMAL) are not present in the meaning extracted from the n-gram features.

4.5.5 Experiment 2: 20 concepts × 26 random n-gram noun features

4.5.5.1 Objective and Methods

In this experiment, we replicate the first experiment's method, with one critical variation - the hand-coded 26 features used by Louwerse (2011), are replaced with 26 random nouns. This is implemented by randomly sub-selecting from a list of 4,554 common nouns³. Then 26 nouns are randomly selected, to create a matrix of 20 verbal descriptors × 26 features, which is analysed using classical MDS. The step of randomly selecting 26 nouns and then generating an MDS plot is repeated 10 times and the average results are used for analysis. In this experiment we are testing the second hypothesis, that *semantic maps and hierarchical clustering will show very weak meaningful associative and taxonomic associations when 26 random noun n-gram features are used*. We find that meaning extraction from language, despite containing sufficient statistical regularities for modelling semantics, in support of symbol interdependency theory, is only sufficiently powerful when language is significantly constrained to increase the symbolic interconnectivity, which the selection of specific features helps enable. Furthermore, the likelihood of the 26 nouns being randomly selected for each iteration to be strongly associated with the 20 verbal descriptors is minimal as we speculate that the noun-to-noun symbol interdependencies to consist of small-world networks, where arbitrary associative links of degree one (first-order co-occurrences) are going to be, on average, information poor.

³ Source: <http://www.desiquintans.com/nounlist>

4.5.5.2 Results

The second experiment's semantic representations, using 26 random noun n-gram features (see *figure 4.3b + e*), are very weak with TREES being the only category that has strong inter-category correlations. These representations are the weakest semantic encodings across all 7 experiments, where associative and taxonomic relations at the superordinate (e.g. ANIMAL vs PLANT) and basic (e.g. PIG vs DOG) levels are the weakest if at all present (TREES being an exception).

4.5.6 Experiment 3: 20 concepts × 26 random n-gram verb features

4.5.6.1 Objective and Methods

This experiment is identical to experiment 2, with one difference - instead of using a subset of 4,554 common nouns, we use 8,537 verbs identified in the University of Colorado's *Unified Verb Index*⁴. Here, the hypothesis is that *semantic maps and hierarchical clustering will show strong meaningful associative and taxonomic associations when 26 random n-gram "verb features" are used*. Verbs are linguistic devices for describing actions, and we therefore predict that the verb-to-noun (all verbal descriptors are nouns) interrelations are going to be more prevalent than in the noun-to-noun scenario. Therefore, more meaningful interrelations between the verbal descriptors are likely to be successfully mediated by verbs. However, we still predict that these associations will not be as taxonomically and associatively strong as in the modeller-selected features in experiment 1. However, if the associations are still sufficiently strong, this would lend even greater support to symbol interdependency hypothesis than experiment 1, because of the more ecological valid modelling approach of not pre-selecting the specific feature words, as this is done randomly. This is a wide-spread problem in cognitive modelling, especially in the domain of semantic cognition, which can blur the

⁴ Source: <http://verbs.colorado.edu/verb-index/>

boundary between whether or not the computationally extracted meaning is a genuine output of the model or biased by the modeller's preconceptions.

4.5.6.2 Results

In experiment 3, using 26 random verb n-gram features, (see figure 4.3c + f), some semantic representations are meaningfully encoded. For example, ROBIN and PENGUIN as well as BIRCH, MAPLE, PINE and OAK are respectively grouped together. Nevertheless, there are some exceptions to this (e.g. ROSE and DOG). The hierarchical cluster analysis however reveals that most basic and in particular superordinate categories are poorly encoded in the semantic representation based on verb n-gram features alone.

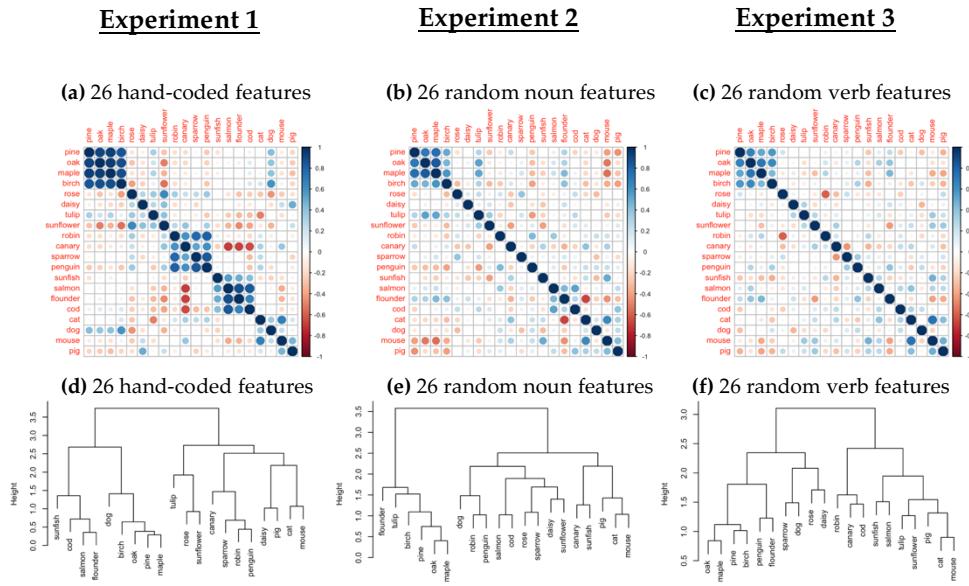


Figure 4.3: The differentiation of conceptual representations based on 26 hand-coded n-gram features (a + d), 26 random noun n-gram features (b + e) and 26 random verb n-gram features (c + f).

4.5.7 Experiment 4: 20 concepts \times 300 / 150 / 50 LSA dimensions

4.5.7.1 Objective and Methods

Here we test LSA's meaning extraction ability by means of the commonly used 300 dimensions, trained on the TASA corpus, as well as two additional variants of 150 and 50 dimensions. We hypothesise that *semantic maps and hierarchical clustering will show the weakest associative and taxonomic associations when 300 LSA dimensions are used.* This is based on our assumption that despite moving towards latent semantic interrelations (e.g. beyond first-order co-occurrences), this additional information will not sufficiently increase the signal-to-noise ratio of linguistic associations. We also hypothesise that *semantic maps and hierarchical clustering will show stronger and more meaningful associative and taxonomic associations when LSA dimensions are gradually reduced (300 \rightarrow 150 \rightarrow 50)*, as the signal-to-noise ratio is likely to increase, leading to superior meaning extraction capabilities, reflected in clearer taxonomic and associative relationships.

4.5.7.2 Results

In the fourth experiment's LSA-300 condition, all semantic relations are the weakest and most stochastic. In the LSA-150 condition, basic categories show some meaningful interrelations (e.g. FISH), while in the superior LSA-50 representation, weaker taxonomic relations start emerging (see *figure 4.4*). These results indicate that using fewer LSA dimensions can improve semantic encoding.

Experiment 4

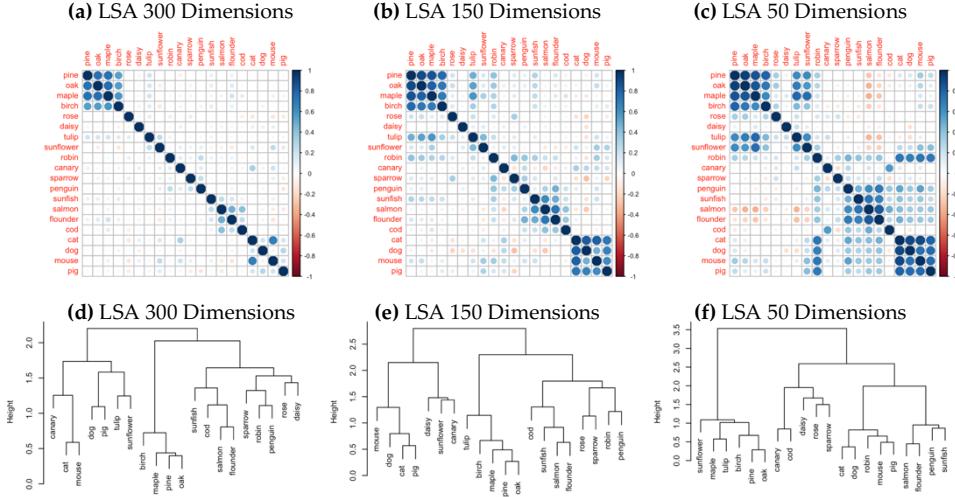


Figure 4.4: The differentiation of conceptual representations based on 300 LSA dimensions (a + d), 150 LSA dimensions (b + e) and 50 LSA dimensions (c + f).

4.5.8 Experiment 5: 20 concepts \times Perceptual Scene Vectors (PSVs)

4.5.8.1 Perceptual Scene Vectors

We operationalise and computationally implement the automatic extraction of *key objects* and scene characteristics from images, for creating *Perceptual Scene Vectors* (PSVs), which we define as a 1D tensor (i.e., a vector) capturing context-specific and cognitively-inspired statistical regularities. Moreover, we also introduce a novel term called *cognitive data representation*, to summarise and justify the underlying cognitive psychological reasons for determining why specific data formats should be used. We define the importance of cognitive data representations by providing an example. Our central aim is to extend Louwerse's symbol interdependency theory beyond language to the broader remit of our continuous streams of experiential information, a large part of which we access through our visual system. However, if we were to suggest that PSVs should be generated based on merely scanning digital photo tags (the keywords people enter to summarise key aspects and/or to categorise their photos), then that would

not be a suitable cognitive data representation for extending Louwerse's theory, given that these tags are linguistic in nature.

Prior to outlining the computational details of PSVs, we clarify two core assumptions underlying our theoretical justification for focusing on (i) only key objects in the surrounding need to be identified and (ii) cognitive data representations are inspired by both brain-based as well as phenomenologically salient features of human experience. Firstly, as discussed earlier, empirical findings (e.g. Intraub et al., 1996) have shown that people have object-level associative frameworks, such that they will be aware of other "things" and "settings" but not (at least explicitly) reference "a blade of grass". Motivated by the concept of basic-level categories (e.g. RABBIT) being more widely used in general contexts, as opposed to superordinate (MAMMAL) or subordinate (LOP) ones, we suggest that PSVs should also capture experiential-level information from real-world scenes. This assumption is important because it determines the particular type of scene segmentation algorithm we use for extracting the cognitively relevant information from the scene. Our motivation to create PSVs is to further the computational study of cognition, as opposed to developing a superior engineering framework. In fact, even the first assumption purposefully leads us to choose a sub-optimal algorithm for semantic applications, as the focus on experiential-level categories inevitably discards a great deal of statistical regularities that could be useful in, for example, web semantic applications.

Secondly, suggesting that PSVs are inspired by both brain-based as well as phenomenologically salient features of human experience does not determine a particular type of representation per se. For example, we strongly argue that cognitive data representations change throughout the early developmental stages of infancy and toddlerhood all the way into early adolescence, gradually becoming more symbolic and less dependent on the immediate surroundings. Biederman (2017) defined five types of characteristics that are important for scene processing. These are as follows:

(i) *support*, (ii) *interposition*, (iii) *probability*, (iv) *position* and (v) *size*. Support refers to the force-dynamic property of most objects in our surroundings being physically supported as opposed to floating in mid-air. Interposition refers to objects mingled together in a visual scene such that based on information from the relative position of one object from another, it is possible to compute who or what is occluding or being occluded. The third type, probability, determines the likelihood of objects co-occurring together. The typical position of objects, *spatial iconicity*, has numerous stereotypical configurations such as roof usually being on top of a house, which has also been a focus area for some EC research (Louwerse, 2008).

Lastly, objects are commonly associated with standard relative sizes. All of these characteristics are cognitively relevant, however, some of them (e.g. support) might be particularly relevant for modelling children's development of object physics. However, in our case of representing the co-occurrences of visual objects and their symbolically associative properties based on context, incorporating all five types of Biederman's categories is unnecessary. We suggest that cognitive data structures should closely match our phenomenological experience of the environment. This means the omission of advanced mathematical or computational strategies merely for optimising algorithmic performance. For example, we do not supplement our PSVs with elaborate *Hierarchical Bayesian classifiers* for assigning a graded set of probabilities for objects that one might typically expect in a given visual scene. Similarly, the phenomenon of *boundary extension* (extrapolating from a limited viewing angle to a wider one) is interpreted as a higher-order associative mental capacity as opposed to one based on lower-level perceptual information grounded in real-world scenes, and so is excluded from the generation of PSVs.

Perceptual Scene Vectors are generated in five steps, always starting from a set of raw photographs⁵ - representing ecologically valid cognitive data representations of the visual world. The first step consists of extracting 10 random images per concept-to-be-modelled. In step 2 we implement Zhao et al.'s (2017) pre-trained *pyramid scene parsing network* (PSPNet), which consists of a specially trained deep learning network with a *pyramid pooling module* (see *figure 4.5*), which is most suited for scene parsing as a result of the global-scene level priors from the pyramid pooling layer integrating contextually and globally relevant information from the visual receptive fields. Zhao and colleagues tested PSPNet against a wide range of other scene parsers using the ADE20K dataset, which includes 20,210 training images and 2,000 validation images⁶, and it consistently performed best-in-class. However, more importantly for our purposes of deriving a suitable cognitive data representation for naturally occurring scenes, PSPNet has the unique ability of specifying the level of *local versus global* level of interest (see *figure 4.6*). At level 1, PSPNet outputs higher-level object masks (e.g. parsing people, tables, computers, cups etc.), whereas at level 2 it extracts object parts (e.g. head, arms, cup handle), and lastly, level 3 extracts parts of object parts (e.g. eyes, nose and mouth). The pre-trained PSPNet outputs a text file with object probabilities for every input image.

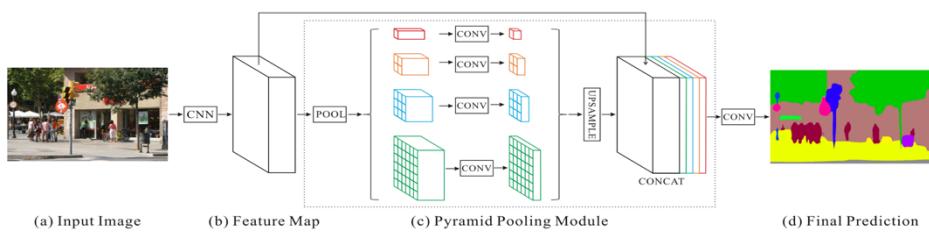


Figure 4.5: Zhao et al.'s PSPNet framework optimised for extracting global-scene-level priors.

⁵ In theory and practice, videos can also be used for increased ecological validity, but we avoid this here due to the intractability of implementing such a computational model on regular CPUs.

⁶ At the time of conducting this study, the test dataset did not exist.

Step 3 extracts all object probabilities based on the first level of abstraction (objects), as this level is considered to be the most relevant format of cognitive data representation (see *figure 4.7*). This creates a smaller text file with only the outputs needed for subsequent stages. At this level of abstraction PSPNet has a *mean intersection over union* (mIOU) accuracy of 85.4% on the scene benchmark dataset PASCAL VOC2012 (Visual Objects Classes). Step 4 transforms the raw text file into a structured data format, where each row corresponds to a particular verbal descriptor in our semantic modelling (e.g. DOG or PINE) and the columns contain the object-level probabilities. For example, in *figure 4.8*, the top-left image depicts a photo of a car including a *building*, that has an object segmentation probability of 73.6%, which is towards the higher end for real-world photographs. The final stage, step 5, binarises the probabilities based on a modeller-defined threshold for filtering out objects/columns automatically identified in the photographs that have probabilities less than 20%. We term this threshold the *attentional filter*, in-part, inspired by Donald Broadbent's (1958) *filter model of selective attention*. Given that cognitive information processing is highly constrained by our limited processing capacity at a given time, this "early filtering out" of information is manifested in our creation of PSVs via the introduction of this attentional filter. Based on these steps, a single PSV for every verbal descriptor being modelled is generated.



Figure 4.6: PSPNet's 3 level of abstraction: (i) objects (level 1), (ii) object parts (level 2), and (iii) parts of object parts (level 3). The network starts with a photograph and outputs a text file containing the probabilities of objects and/or object parts.

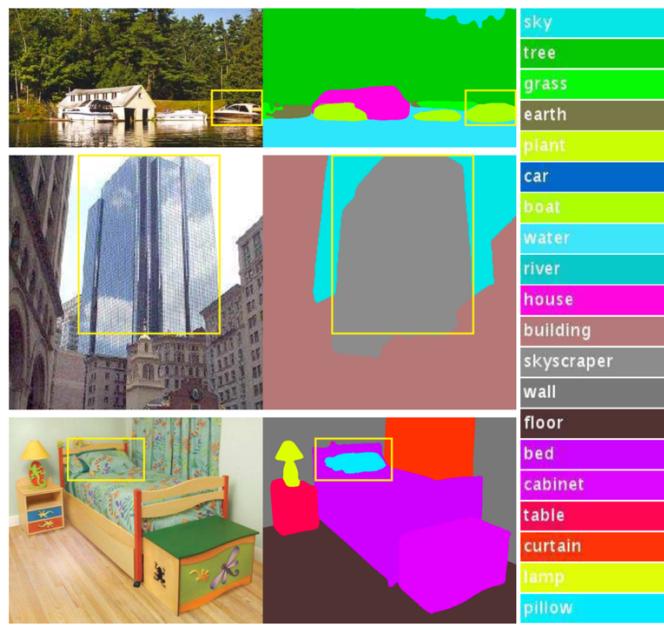


Figure 4.7: Example from Zhou et al. demonstrating object-level scene segmentation.

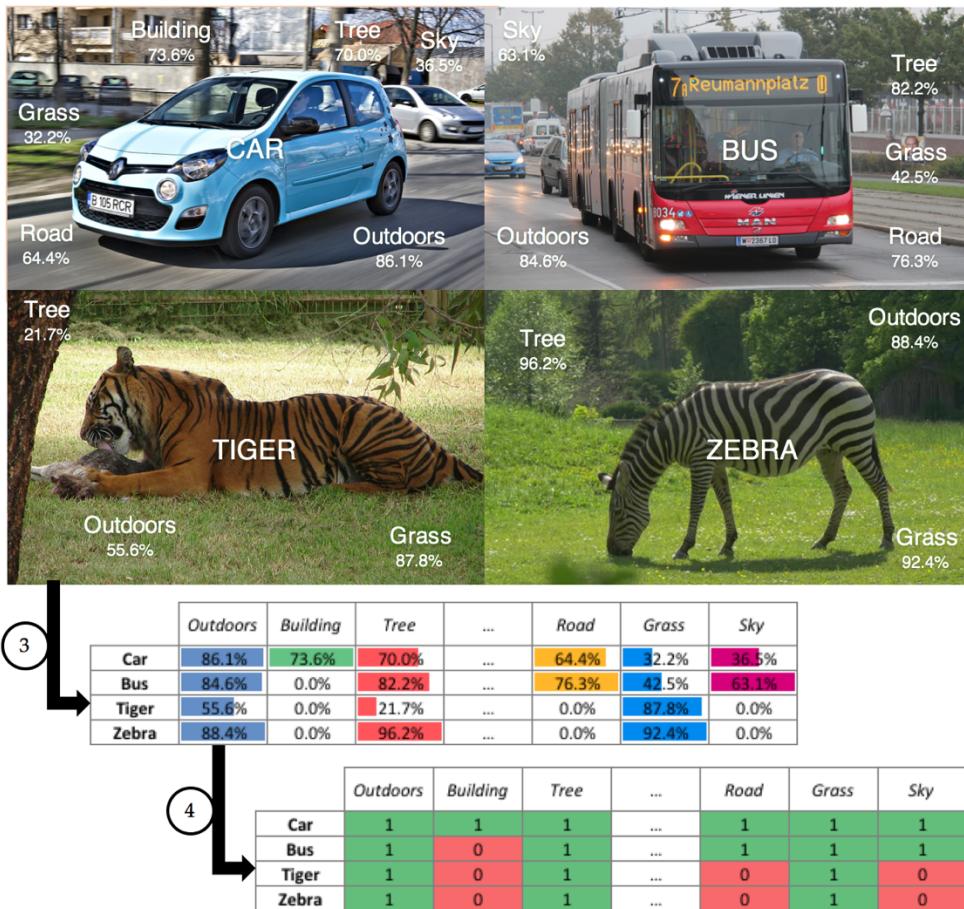


Figure 4.8: Schematic example of extracting the probabilities and binarising.

4.5.8.2 Objective and Methods

The methodology for this experiment differs from experiments 1 - 4. This is the first experiment where neither language-based inputs (first-order word co-occurrences/LSA dimensions) nor modeller-selected features are used for modelling semantics. Instead, 20 PSVs (one per verbal descriptor) are the inputs to a simple feedforward neural network. However, for this particular experiment, we limit the number of PSV bits per verbal descriptor to 26 in order to match the informational capacity of each verbal descriptor representation to minimise biases in comparing with results from experiments 1 - 4. This limit is implemented by first selecting the binarised PSV properties (across all verbal descriptors) ranked by the proportion of average active bits across all 20 verbal descriptors, although duplicates are omitted based on the order. For example, if one PSV duplicate is *grass* = {1,1,1,1,...,0,0,0,0} and another *earth* = {1,1,1,1,...,0,0,0,0}, then *earth* is omitted. However, this is only a necessity for balancing the need of having 26 bits to have comparability with the other experiments.

The 26 PSVs are the 26 inputs to the neural network, with 20 hidden layer units, and 20 outputs, one-hot encoding of the verbal descriptors. The ReLU (rectified linear units) activation function is used, which is the simplest activation function one can use and is: $f(x) = \max(x, 0)$; the output is equal to the input as long as it is greater than zero. A dropout layer is used with a rate of 0.4 to avoid overfitting. A final softmax layer is used for transforming the 20-bit output to sum to 1, converting it to a probability. The network's learning rate is set to a constant 0.001, and the network is trained for 500 epochs to a classification accuracy of 100%. Once the neural network training is completed, the inputs are fed back into the network, whilst the hidden layer representations for each PSV is recorded. These hidden-layer representations are then analysed using classical MDS.

We predict that the *semantic maps and hierarchical clustering will show the most meaningful associative and taxonomic associations when automatically extracted Perceptual Scene Vectors (PSVs) are used*. Support for this hypothesis

would provide a strong argument in favour of extending the symbol interdependency hypothesis beyond language, to real-world scenes.

4.5.8.3 Results

In our fifth and crucial experiment, which tests our novel *Perceptual Scene Vectors* (PSVs) approach for automatic encoding of semantic associations without the need for hand-coded features, the results show the presence of the strongest taxonomic and associative conceptual representations (see *figure 4.9*). In this experiment, we see clear evidence of superordinate classification between ANIMALS and PLANTS (revealed by patterns in the corrplot), and also stark basic-level distinctions between categories. The only “anomalous” concept leaf in the hierarchical cluster tree, is PENGUIN, which is yet meaningfully grouped together with FISH, although being slightly further away from the more prototypical members, in-line with our expectations. Similarly, the concept CANARY (a common pet), despite being grouped together with other BIRDS, is also slightly differentiated. Even the more nuanced distinctions, e.g. ROSE and TULIP versus DAISY and SUNFLOWER or SALMON and COD (food) versus SUNFISH and FLOUNDER, are structured as one would expect, which we discuss in the next section, as it is based on a unique advantage of PSVs compared to feature-based methods. The absolute and relative heights (Euclidean distances) between all classes and subclasses are clearly differentiated, and conceptual grouping do not suffer from spurious correlations, as seen in the representations from experiments 1 through 4.

Experiment 5

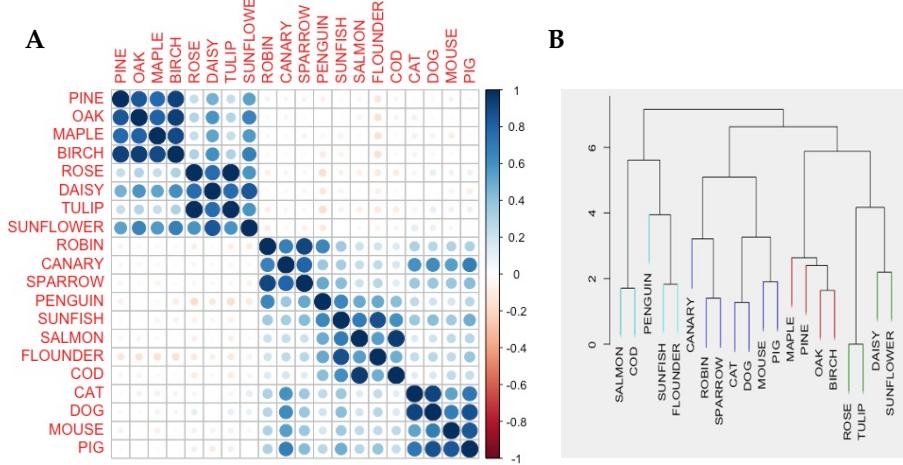


Figure 4.9: Conceptual representations of 20 verbal descriptors with discernible taxonomic and associative hierarchies based on the hidden layer associations of the neural network trained using PSVs. (A) The CorrPlot reveals robust categorisation of trees, flowers, birds, fish and mammals as well as higher-order distinctions between animals and plants. (B) The Hclust further supports strong discriminations between the 20 concepts, which is an emergent property grounded in the statistical regularities of real-world scenes.

4.5.9 Experiment 6: 20 new concepts \times 300 LSA dimensions

4.5.9.1 Objective and Methods

This is a partial replication of experiment 4, but with only the standard 300 LSA dimensions and 20 new verbal descriptors (see table 4.1). The objective is to have an additional set of results for benchmarking the results from experiment 7. Similar to our hypothesis for experiment 4, we once again hypothesise that *semantic maps and hierarchical clustering will show weak and random associative and taxonomic associations when only LSA's 300 dimensions are used*.

4.5.9.2 Results

In our final two experiments, we directly compare the standard LSA 300 dimensions (experiment 6) and the PSVs (experiment 7) using a new set of 20 verbal descriptors, see figure 4.10. The results of experiment 6 reveal

the weakest associative and taxonomic relationships, and an abundance of spurious weak correlations, even though more sensible associations also appear (e.g. BUTTER, JAM and TOAST). Visually inspecting the Hclust tree for experiment 6, reveals a common trait of LSA-300 dimensions (experiments 4a and 6) for encoding conceptual representations - the lack of coherent associations leading to a very flat and undifferentiated tree. The corrplot of LSA-300 dimensions also reflects this with the presence of a dominant diagonal “activation line”, where the verbal descriptors appear to be almost independent of one another.

4.5.10 Experiment 7: 20 new concepts × Perceptual Scene Vectors (PSVs)

4.5.10.1 Objective and Methods

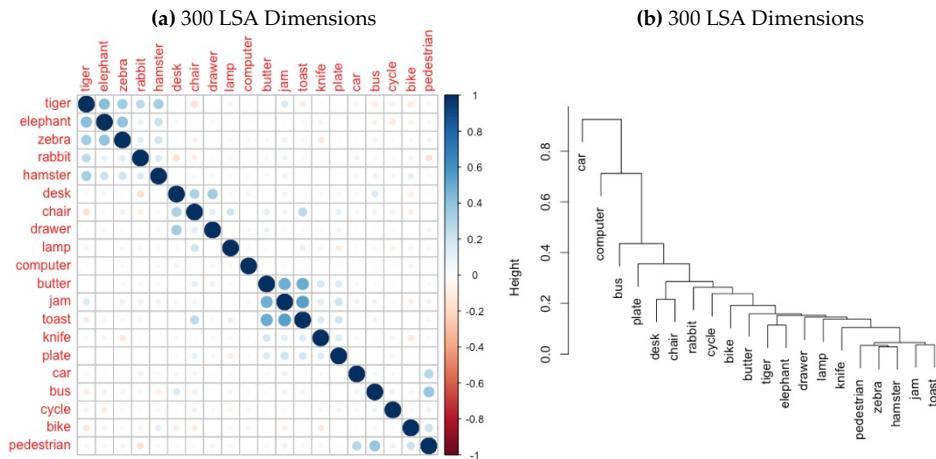
This is a partial replication of experiment 5 but using the 20 new verbal descriptors. The methods and hypotheses are identical to those of experiment 5.

4.5.10.2 Results

In contrast to the results from experiment 6, in experiment 7, the same verbal descriptors modelled using PSVs show very strong associative relations, where ANIMALS, OFFICE EQUIPMENT, FOOD & KITCHENWARE and TRANSPORTATION are neatly grouped together in their respective categories. Evidence supporting taxonomic differentiation is limited, with the exception of SMALL and BIG ANIMALS as well as CYCLE and BIKE versus TRAIN, CAR and BUS being grouped independently. Superordinate-level distinctions between ANIMATE and INANIMATE relations seems to appear from the Hclust, as it is the first divergence point, but the corrplot does not provide sufficient differentiation to justify this interpretation. The experimental design

limitation of not having sufficient taxonomic variability in these new verbal descriptors is covered in our discussion.

Experiment 6



Experiment 7

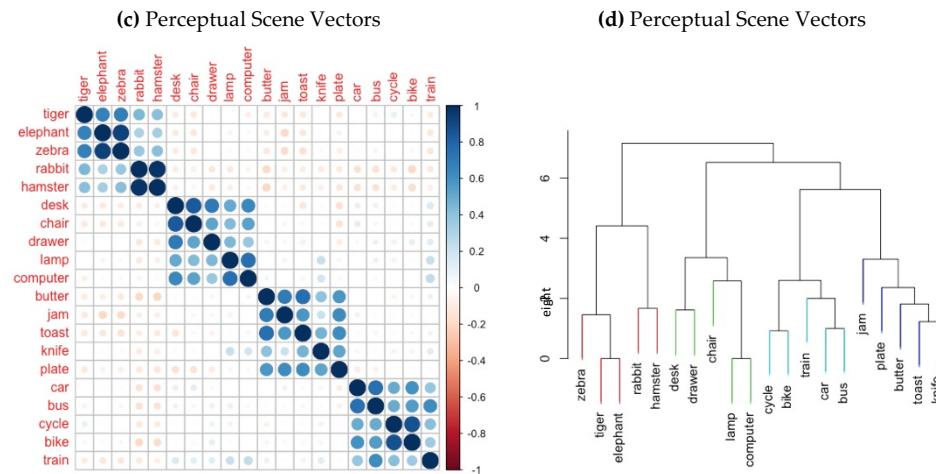


Figure 4.10: (a + b) Conceptual representations of 20 novel animate and inanimate verbal descriptors based on 300 LSA-dimensions. (c + d) The relationships of the same 20 concepts based on the hidden layer associations of the simple feedforward neural network trained using PSVs. By comparing the inter-correlations of the conceptual representations generated by the LSA algorithm (a) and PSVs technique (c), we illustrate the superior associative relations between and within the animal, office, kitchen and transportation categories. Moreover, when comparing the two hierarchical clustering solutions, the taxonomic associations based on PSVs (d) is more discriminating than those derived through LSA (b). Animate and inanimate differences in the conceptual representations is also the first point of differentiation when using PSVs as can be seen by the first tree split in plot (d).

4.5.11 Quantifying Semantic Coherence

We outline a range of nuanced hypotheses for the 7 experiments in this chapter, where we make qualitative predictions about the extraction of meaningful representations that have strong associative and taxonomic relationships (see *table 4.2*).

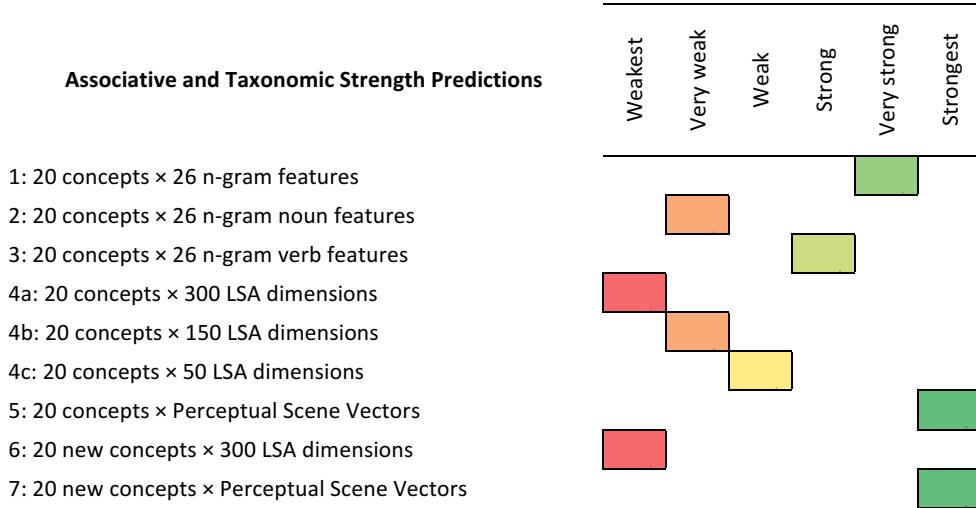


Table 4.2: Summary of our predictions for experiments 1 through 7. Weak predictions indicate very poor associative and taxonomic relationships, whereas strong predictions more meaningful relationships.

Creating a continuum, in our case, from the weakest to strongest relationship, helps provide an ordinal structure to our hypothesis dimension, for distinguishing between language- and scene-based meaning spaces. Furthermore, it allows us to explore more quantitative metrics that compare how well an algorithm can extract meaning.

Semantic representations, especially in the field of cognitive modelling, are typically evaluated using statistical methods such as MDS, principal component analysis (PCA) and/or hierarchical clustering (Hclust). Our focus here will be on hierarchical clustering due to the prevalence of this particular statistical technique for the qualitative evaluation of semantic model outputs across a range of empirical and modelling investigations (Small et al., 1995; Kriegeskorte et al., 2008; Rogers

& McClelland, 2004). Agglomerative hierarchical clustering algorithms typically output a tree diagram/dendrogram depicting the cluster membership and taxonomy. Agglomerative hierarchical clustering specifically starts with all variables as separate clusters and then iteratively joining the two most similar clusters together until all clusters are grouped together. However, once clusters are grouped together, they are not split out during subsequent iterations. The vertical distance (usually Euclidean or Jaccard) is a measure of dissimilarity, with greater distances representing larger differences between the groups.

Traditionally, hierarchical clustering outputs take the form of a tree diagram and are evaluated based on whether or not they qualitatively “make sense”, essentially reducing this to an evaluation of whether or not common-sense associative and taxonomic relationships are meaningfully captured by the model. For example, are *animate* and *inanimate* objects subdivided into different groups? Equally, our intuitions allow us to understand that concepts such as PENGUINS could sensibly fall in either the bird or the fish category but not in the inanimate cluster. When comparing more than one hierarchical cluster analysis based on n-dimensional multivariate data, such intuitions may well be sufficient when n is a relatively small number like in most cognitive modelling “toy datasets” of typically no more than 20 concepts. However, a robust quantitative measure of semantic similarity is pertinent for making reliable comparisons across a range of different models, which is currently lacking in the contemporary cognitive semantics literature.

In the extant literature, the need to quantify the semantic association plots also largely remains elusive, although the internal validity of clusters is typically analysed by visually inspecting the rate of decay of the mean square error (MSE) as the number of clusters increase (similar to the screeplot in *factor analysis*). Ball and Hall (1965) are probably the first to review the use of MSEs for obtaining the optimal number of clusters. Typically, there is an “elbow-point”, which determines that after k clusters,

additional clusters are extracted with increasing levels of diminishing returns, mathematically captured by the *proportion of variance explained* (see Halkidi, Batistakis, & Vazirgiannis, 2002). However, *extrinsic validity* of clusters, comparing two cluster solutions against one another or in relation to an independent *ground truth* (e.g. data collected from psychological studies) is an underexplored area within semantic cognition research. A notable exception is Pulvermüller et al.'s (2010) brain mapping study.

Pulvermüller and colleagues investigated the connections within the frontotemporal cortex to discriminate distributed lexical and category-specific networks, using *k-means*, a commonly used clustering algorithm for partitioning the data into k clusters, where k is determined by the analyst. Pulvermüller et al. compare the clustering similarity across experiments using the *Rand index* (RI), which calculates the *degree of similarity* between two sets of clusters by comparing the class labels, and ranges from 0 to 1, where 1 indicates perfect correspondence between two classifications (Rand, 1971). However, we recommend using the *adjusted Rand index* (ARI) for cognitive modelling purposes, which is the *corrected-for-chance* form of the original Rand index (Hubert & Arabie, 1985). Pulvermüller et al. analyse neuroimaging data, where semantic networks are determined based on the co-activation of thousands of voxels, and the probability of cluster labels overlapping with one another is very small. However, much of cognitive semantics research, including the experiments in this chapter, consist of toy models with extremely small datasets, and therefore chance-level overlaps across vectors of cluster membership are more likely to occur, and the difference between the RI and ARI is likely to be significantly larger for smaller datasets, making the ARI a more suitable metric for comparing cluster similarities between membership groups.

	Adjusted Rand Index
1: 20 concepts × 26 n-gram features	0.79
2: 20 concepts × 26 n-gram noun features	0.44
3: 20 concepts × 26 n-gram verb features	0.51
4a: 20 concepts × 300 LSA dimensions	0.44
4b: 20 concepts × 150 LSA dimensions	0.59
4c: 20 concepts × 50 LSA dimensions	0.64
5: 20 concepts × Perceptual Scene Vectors	0.93
6: 20 new concepts × 300 LSA dimensions	0.41
7: 20 new concepts × Perceptual Scene Vectors	0.88

Table 4.3: Summary of Adjusted Rand Index (ARI) across all 7 experiments, quantifying the similarity in semantic clustering between the various language and Perceptual Scene Vector (PSV) experiments and the ground truth clustering from Rogers and McClelland (2004).

Finally, we quantify the quality of our semantic modelling representations by comparing them to an independent *ground truth* - a test for external validity in our hierarchical clusters. In experiments 1 - 5, this ground truth is the original Rogers and McClelland (2004) connectionist feature-based network model, while for experiments 6 and 7, inspired by Rogers and McClelland's model, we create a hand-coded feature-based model. Our *Adjusted Rand Index* (ARI) is calculated by averaging independent ARIs over several cuts of the tree (range = 2, 4, 6, 8) for all 7 experiments' Hclust membership vectors versus their respective ground truth representations' Hclust vectors. The results are summarised in *table 4.3*. Generally, across all the experiments, we can see that semantic encodings based on PSVs perform the best (experiments 5 & 7), with the highest ARI score being 0.93, indicating an extremely high level of correspondence with the ground truth representation. This is closely followed by Louwerse's (2011) conceptual replication using 26 LSA feature (experiment 1). The representations with the lowest ARI scores are, the standard LSA 300 dimensions (experiments 4a & 6), conditions with 26 random noun n-gram features (experiment 2), 26 random verb n-gram features (experiment 3). The representation of experiment 6, with the 20 new verbal descriptors and 300 LSA dimensions, has the lowest ARI score

of 0.41. However, experiments 4b and 4c, where the LSA dimensions are respectively reduced to the first 150 and 50 dimensions have ARI scores in the midrange between 0.59 and 0.64.

4.6 Discussion

The present chapter's motivation started with the conceptual replication of Louwerse's (2011) results on extracting meaning from surface-level statistical regularities in language but with the aim of extending *symbol interdependency* beyond language. Our extension of symbol interdependency, in this chapter, focused on grounding it in real-world scenes, based on ecologically valid cognitive data representations, a framework we introduce for moving beyond simple toy datasets and modeller-determined lists of features.

We successfully replicated Louwerse's finding of using surface-level language structures for capturing sufficient symbol interdependencies for extracting semantic representations with many of the properties found in traditional connectionist models of semantic cognition. In doing so, we independently provide support for Louwerse's symbol interdependency hypothesis, and also support the argument that the "input data" (e.g. language) is more important than the algorithm itself for meaning extraction.

Both a qualitative analysis of the representations and a quantitative estimation of the correspondence between our language-based models and the original Rogers and McClelland (2004) model using the Adjusted Rand Index provide strong evidence in support of statistical regularities found in surface-level language patterns. To our best knowledge, this is the first time semantic representations have been quantitatively evaluated to test for external validity using an independent *ground truth*. Nevertheless, one general limitation of much corpus-based semantic modelling, including our LSA-based experiments, is the comparability of LSA spaces. Even though

we use the same corpus as Louwerse (TASA), LSA spaces are typically calculated with different default parameters. Therefore, some of the smaller differences we have noticed in the taxonomic associations between our model and that of Louwerse's might be due to these "hidden parameters" used for generating our respective LSA spaces. However, our findings still robustly support Louwerse's original conclusions. To overcome this drawback, we recommend that the cognitive modelling community should create a set of *standard benchmark spaces* that would provide a robust testbed for all researchers. Modelling with a set of pre-determined features obfuscates the finding of computationally extracting meaningful semantic representations from first-order linguistic co-occurrences as the outcome could be influenced by the modeller's knowledge of particular features that contain sufficient relationships with the verbal descriptors being modelled in the first place.

The core theoretical objective of experiments 2 (26 random noun n-gram features) and 3 (26 random verb n-gram features) is centred on avoiding modeller-defined features. We reason that this eliminates the limitation of *information leakage* from the modeller to the semantic representation generated computationally. As we predicted, when using 26 random noun-based features, the semantic representations are very weak. In fact, the representations are so weak that much of the variability in the correlations is largely spurious. Where stronger correlations do exist (e.g. TREES), an informal analysis of the n-grams of the words PINE, OAK, MAPLE and BIRCH reveals, unsurprisingly, the prevalence of the word "tree" in the TASA corpus, which explains the more robust associative relations between trees even when other relations are stochastic. In our third experiment, with random verb features, our prediction was only directionally correct, as we expected to find even stronger semantic associations than we did. The use of verb features led to a slightly superior semantic representation when compared to noun features, which we think can be largely attributed to verbs being more useful for distinguishing

animate concepts from one another and from inanimate ones, although this remains a speculative interpretation.

In our standard LSA 300-dimensions model (experiment 4), we find the weakest semantic associations, which is surprising given the wide range of LSA applications. LSA's 300 dimensions are ordered from most to least important in terms of proportion of corpus variance explained. Interestingly, we find evidence for a smaller number of stronger dimensions being superior for meaning extraction, because LSA 50 dimensions ($ARI = 0.64$) are quantitatively and qualitatively superior than 150 dimensions ($ARI = 0.59$), which in turn is considerably more conceptually meaningful than 300 dimensions ($ARI = 0.44$). In our case of concrete verbal descriptors, we have evidence for fewer dimensions being superior than more dimensions, even though that is unlikely to generalise for a wide range of concepts that LSA typically has to model. Therefore, given this finding, we are certainly not recommending that fewer LSA dimensions are better for meaning extraction *in general*.

Landauer and Dumais (1997) originally recommended the *semantic granularity* of the LSA algorithm to be set to 300 dimensions, which was found to be optimal for a range of different corpus-based analyses. However, the cognitive scientists Griffiths, Steyvers and Tenenbaum (2007) tested 100 to 700 dimensions (intervals of 100) and found 500 dimensions to be optimal. One key reason for the difference between our results and those of Griffiths et al. might be due to the fact that their study examines *topic models* across thousands of complex semantic networks (on a cut-down version of TASA) where additional LSA dimensions are more likely to contribute "semantic signal", whereas in our small dataset of simple verbal descriptors, additional dimensions are likely to merely increase "semantic noise". Therefore, we do not make recommendations on the optimal number of LSA dimensions, as this is not the objective of experiment 4. Our focus is on the nature of the semantic representations as a function of number of LSA dimensions.

In our experiments, both the reduction of LSA dimensions and the hand-coding of features lead to more meaningful semantic spaces. Based on experiments 1 through 4, we suggest that a common advantage exploited by both the reduction of LSA dimensions and the hand-coding of features is the increase in the signal-to-noise ratio, facilitating computational meaning extraction. The reduction of LSA dimensions allows to filter out less relevant/noisy covariation patterns from the latter dimensions, which can interfere and distort the earlier and more relevant latent semantic dimensions. In the case of modeller-selected n-gram features, we, as the modellers, are using our own cognitive apparatus, to instil the artificial cognitive model with the most meaningful characteristics available to us. This helps encode the most relevant surface-level covariations for describing our verbal descriptors. For example, when a computational modeller decides to use a specific feature such as BARK, this significantly reduces the computational challenge of surfacing the relevant linguistic surface structures in the first place. Moreover, from a cognitive modelling perspective, it does not provide sufficient mechanistic details on meaning extraction, even though, from a linguistic standpoint, it offers strong support for a language-centred view on conceptual processing. In other words, we argue, the modeller is engaged in arguably the most complex and difficult phase of meaning extraction - relevant feature extraction, which should instead be mechanistically accounted for in the computational model itself.

The *relevance problem* is a core theoretical issue for cognitive modelling; because of the multitude of possible relations, it is very difficult to efficiently determine which ones are relevant (Blasch et al., 2006). Within the domain of conceptual processing this is perhaps an even greater problem, not dissimilar to the classic *frame problem* in AI which was first outlined by McCarthy and Hayes (1969) and formally outlines the difficulty for artificial systems to be able to provide rational default assumptions, based on environmental constraints. Human and non-human animals

understand relevance typically based on prior experience or instincts, unless they find themselves in particularly novel situations. However, for the development of computational intelligence, more broadly, and computational cognitive models, in particular, this is a difficult challenge.

In two of our fundamental experiments (5 and 7), we demonstrate the computational viability and advantage of using *Perceptual Scene Vectors* (PSVs), our original computational contribution to the study of modelling semantic cognition. Using the pre-trained general-purpose *PSPNet* deep learning model (Zhao et al., 2017), we automatically extract object-level associations (not features), which across numerous images (only 10 per verbal descriptor) generate a little dataset containing rich object-to-object statistical co-variations. Using these ecologically valid and automatically extracted features grounded in our naturally occurring visual experiences, we model a simple feedforward neural network. Thus, the richness in our semantic representations does not rely on complex algorithms or big data. The hidden layer representations of two separate small training datasets (experiments 5 and 7) are associatively and taxonomically more meaningful and have the best external validity as measured by the correspondences with their respective ground truths. Comparing the results of experiments with PSVs with those using language-based inputs further supports the feasibility and power of PSVs. PSVs permit the querying of why associations or taxonomies are present in resultant semantic spaces, in non-trivial ways, in other words, beyond the antecedent of “the modeller hand-coded this specification”. For example, in experiment 5’s PSV-based semantic space, CANARY is in the bird class but still differentiates from ROBIN and SPARROW. Comparing the PSVs of the three birds revealed that amongst birds, CANARY uniquely contains *cage*, as is also the case for RABBIT and HAMSTER in experiment 7, due to their role as pets. Thus, grounded-representations are not only ecologically valid cognitive modelling inputs, but can also help reveal nuanced relations omitted by hand-coded features.

In this chapter we have also found a novel application for an existing quantitative metric for determining the external validity of clusters, in our case the outputs of hierarchical cluster analyses. We propose that the use of the Adjusted Rand Index (ARI) is most suited to small-scale datasets typically used in cognitive modelling of conceptual representations as it corrects for chance-level cluster membership overlaps which are more likely to disproportionately bias the comparison of smaller sets of semantic representations. Our tiered approach of cutting the hierarchical tree at several levels and generating a mean ARI for each tree should be seen as a simple comparison across superordinate, basic and subordinate concept representations. One limitation of our aggregation strategy is that we are giving multiple levels of the hierarchical cluster tree the same *semantic importance*. So a misclassification at the superordinate level and the basic level are equally important, which is not cognitively plausible. Although, in future work, we could create a weighted mean based on empirically derived importance factors. Additionally, we would include verbal descriptors spanning all three levels of concept categories given that we did not have subordinate level concept discriminations (e.g. SPORTS CAR versus HATCHBACK).

In contrast to the original verbal descriptors used by Louwerse (2011), one of the key limitations of our new set of verbal descriptors is the lack of naturally occurring taxonomic structure. Although topic-level associations are present, higher-order pan-topic and taxonomic relations remain elusive, which limits our ability to investigate the semantic representations in more detail. A second, and potentially more severe limitation, is the approach to our external validation. In our experiments, we compare our model outputs with those of other models which consisted of the Rogers and McClelland (2004) model and another feature-based model that we have developed ourselves. Clearly, this is not ideal for evaluating computational semantic spaces objectively. In the latter case (experiment 7), information leakage from the present author could have

easily inflated the Adjusted Rand Index unconsciously, because we are comparing our PSV-based model against a feature-based model we developed. In our analyses, the traditional feature-based models provide a convenient testbed for quantification purposes. However, our recommendation for future evaluations of external validity is to collect empirical data from human participants and treat that as the ground truth for genuine semantic representations. Ontologically, this would obviously be preferable given the aim of cognitive science is to better understand and model actual human cognition.

In closing, we summarise that symbol interdependency is not limited to the latent and surface semantic associations in language, and that our busy visual world contains rich statistical regularities, from which high-quality meaning spaces can be computationally extracted. This automatic mining of meaning grounded in real-world scenes also helps address the relevance problem in cognitive science and artificial intelligence by providing an ecologically valid reference frame, our PSV, and simultaneously circumvents the dilemma of information leakage from the cognitive modeller.

Chapter 5

Grounding Concrete versus Abstract Semantics

5.1 Abstract

Empirical and computational studies on semantics are usually limited to concrete concepts, despite the importance and prevalence of abstract words in the human lexicon. Recently, there has been an increased focus on describing the content of abstract concepts through *introspective*, *emotional*, *metaphorical* and *situational* descriptions (Borghi et al., 2017). The literature converges on *emotions* being essential for abstract words, hypothesised as *embodied abstract semantics* (Kousta et al., 2011). We first replicate the *concreteness continuum* by re-analysing data from a large-scale normative study (Brysbaert et al., 2014). Then, across three novel computational experiments, we compare situationally grounded concepts with traditional language-based representations. After externally validating PSVs using a neuroimaging benchmark, we find PSVs can only successfully represent more concrete words. Lastly, we develop our new

scene2vec representation, by extending PSVs with emotion expressions extracted from photographs which yield noticeably enriched semantic representations across the concreteness spectrum, despite a lower performance for more abstract concepts. Our original contribution of modelling semantics using emotions only partially supports the *embodied abstract semantics* hypothesis and indicates that there is more to representing abstract meaning than emotions alone.

5.2 Introduction

The English language contains a significant proportion of both concrete and abstract words (Kousta et al., 2011); even though concrete words appear more frequently, the vast majority of words in the English lexicon are abstract (Recchia & Jones, 2012). Both cognitive science and artificial intelligence have principally focused on studying concepts from prototypical concrete categories like *birds*, *vehicles* and *artefacts*. This emphasis is understandable given that these are semantic entities with both a mental representation and a physical instantiation and are easier to operationalise in empirical and computational studies. Moreover, research has shown that concrete words are more accessible to recall than abstract words (Walker & Hulme, 1999), which might account for a methodological bias towards the selection of concrete words during empirical and computational experimental designs. Abstract concepts (e.g. *justice*, *calculus*) compared to concrete concepts (e.g. *cat*, *laptop*) also lack delineated referents but are more likely to be used across a broader range of semantic contexts. Therefore, we argue, the *concept-to-concept* associations for abstract words can be weaker due to a lack of referents; yet, more of these weaker associations for abstract concepts can be flexibly incorporated by new relational inferences - a key facet of higher-order cognition - *combinatorial generalisation*. However, the first step to understanding this combinatorial constitution of abstract concepts is to investigate the

semantic content of abstract words. A mechanistic account of how this content is learnt from real-world experiences and then represented in a semantic memory system will further our understanding of the interplay between concrete and abstract concepts and their role in general human cognition. Therefore, abstract concepts provide a crucial litmus test for any general theory of cognitive semantics as well as more specific theoretical perspectives such as grounded cognition. However, we first need to understand *why* abstract and concrete concepts are different in the first place and then to determine whether or not they are genuinely dichotomous.

5.2.1 Comparing Concrete and Abstract Words

There are currently many perspectives on the similarities and differences between concrete and abstract words, and a thorough review of this exceeds our present scope. However, the most relevant position to our grounded cognition stance originates from Barsalou and Wiemer-Hastings' (2005) exploratory research. Barsalou and Wiemer-Hastings investigated how abstract concepts could be situated using a qualitative-quantitative framework. Participants consisted of a small group of 24 undergraduates, who were asked to produce characteristic properties of three *abstract* (*truth, freedom, invention*), *intermediate* (*cooking, farming, carpeting*) and *concrete* (*bird, car, sofa*) concepts. Participants reported the thoughts that came to mind and described the *typical characteristics* of the different words. Each participant had one minute per word and was prompted with "please continue to describe your thoughts as they come to mind" (p.139) after five seconds of pausing. From this qualitative research, combined with detailed quantitative analysis of the participants' verbalisations (video recorded), using *coding schemes* (Wu & Barsalou, 2009), the study was able to establish the relative proportions of taxonomic, object-specific, setting/event and introspective contents present in the self-generated verbalisations. They found support for three out of four of their hypotheses. First, both concrete

and abstract concepts share situational content, because participants frequently referred to different actions, events, goals and affective reactions associated with settings relevant to a given concept. However, this is particularly true for abstract words, which contradicts Schwanenflugel's (1991) *context availability theory*, in which abstract concepts are predicted to have weaker connections with situational information.

Second, Barsalou and Wiemer-Hastings (2005) also found that even though both concrete and abstract concepts share situational content, the specific focus of the content differs. Abstract concepts are unique in that their *situational focus* is not only distributed across a range of complex situational configurations but also is grounded in more *introspective* content constituting most prominently *emotions*, followed by *social* and *event* content and less importantly *physical settings*. However, for concrete concepts, the focus is more on the specific *object-to-object* interrelations as well as the *background* contexts. Relatedly, in a property verification task, Wiemer-Hastings and Xu (2005) found that introspective features were more important for accounting for abstract words. Their third hypothesis, somewhat related to their second hypothesis, claimed that abstract concepts are more complex than concrete concepts due to the distributed nature of the situations for abstract concepts, which was also supported. The only hypothesis for which they did not find any supporting evidence for is that the content of abstract concepts is simulated, although this would be challenging to determine from an exploratory qualitative⁷ study. This study provides evidence for some universality in the mechanistic account underlying semantic representations of both concrete and abstract words. Although Barsalou and Wiemer-Hastings (2005) do not discuss the mechanistic implications of their exploratory findings, we believe that the commonalities in situational information across concepts suggest

⁷ Technically speaking, Barsalou and Wiemer-Hastings (2005) was qualitative in their data collection phase although their analysis consisted of both qualitative (e.g. coding schemes) and quantitative techniques (e.g. ANOVAs).

significant overlaps in concept representations irrespective of concreteness. There are also likely to be mechanistically distinct processes underlying concrete and abstract concepts such as the differential role played by introspective information. Next, we turn our attention to such mechanistic conceptualisations of abstract words.

5.2.2 Modelling Abstract Semantics

Only very recently, have some researchers started incorporating abstract words in their computational models. For example, Hoffman et al. (2018) recently extended the connectionist hub-and-spoke account of semantic cognition to apply it to abstract words. At its core, Hoffman et al.'s modelling relies on co-occurrence patterns within a linguistic distributional system, *supervised topic models*. This iteration of the hub-and-spoke model also incorporates cognitive control and the *linguistic context* shaping meaning based on linguistic and non-linguistic data, which they claim is *embodied*. We disagree with this claim given the non-ecologically valid stimuli used in their computational experiments. Despite Hoffman et al.'s aims of reconciling disembodied and embodied perspectives and including both concrete and abstract words, their model implicitly assumes a bipolar distinction between concrete and abstract concepts given the sharp contrast of stimuli representations based on either *pure embodiment* or *derived embodiment*. Hoffman and colleagues promote a more direct integration of associations between verbal and non-verbal properties ("embodied view") and the computational linguistic perspective in favour of decoding meaning from word distributions across language (disembodied view). Given the relevance of Hoffman et al.'s research to our current focus on computationally grounding concrete and abstract concepts, we provide a detailed review of their work. This is followed by a review of the critical differences in the theoretical assumptions of Hoffman et al. compared to our present research. We then develop an alternative approach, without

reliance on distributed linguistic representations, which is a natural extension of our work on perceptual scene vectors (chapter 4).

Hoffman et al.'s (2018) computational model uses a *recurrent neural network* with 590 units, subdivided into five groups (one central hub and four spokes), with 64 *verbal input* and 64 *verbal prediction* units (output) comprising the concept labels. The 64 concepts consist of 22 concrete concepts, 32 abstract concepts, and ten homonyms. In the architecture, 162 units encode the "sensorimotor properties", while the *context* and *amodal* hub each have 150 units (see *figure 5.1*). Given the recurrent nature of the model, the hub representation is connected bidirectionally with the sensorimotor properties, verbal predictions and context units, while the verbal inputs are unidirectional, feeding into the amodal hub representation. This allows the model to learn sequences of words, such that the preceding word can influence the conceptualisation of the upcoming word. Both the hub and the context pools are hidden layer units, thus are based on gradually learnt associations. The authors also argue that the "recurrent architecture allows the model to develop representations that are sensitive to context" (p. 14).

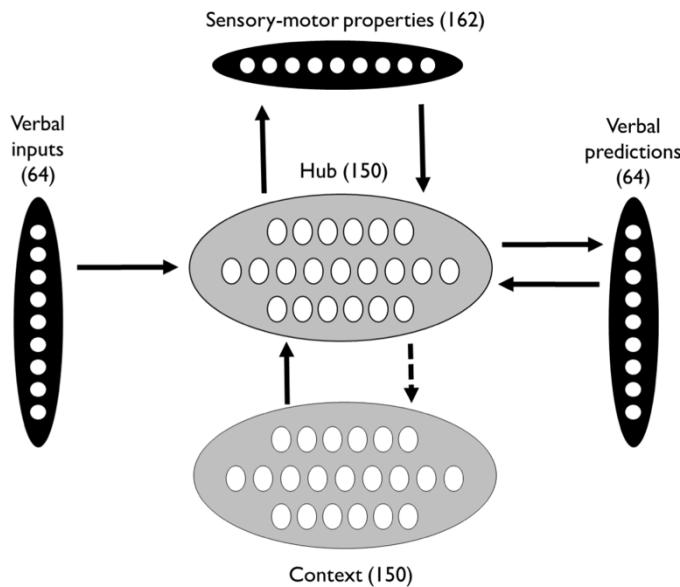


Figure 5.1: The recurrent neural network from Hoffmann et al. (2018).

Hoffman and colleagues use a highly simplified “artificial environment” in which they hand-code sensorimotor properties by creating six overlapping and three unique bit-representations for category neighbours. Therefore, all the stimuli used in their study are created based on *a priori* assumptions on the nature of concrete and abstract concepts. Even though we have been critical of such hand-coding practices, from a grounded cognition perspective, Hoffman et al. do state that their main interest is to implement the most relevant aspect of sensorimotor properties, which is that category neighbours typically share more of these properties than concepts from different categories. Finally, Hoffman et al. generate a synthetic training corpus of 400,000 episodes for training the model across a range of different contexts. The training episodes are generated using a pseudo-stochastic combination of linguistically-derived topic models (for details, see Griffiths et al., 2007) and either constant sensorimotor properties or verbal inputs depending on the specific simulation run across a series of experiments. Hoffman and colleagues argue that their hybrid model is based on cognitive neuroscience evidence that concepts are grounded in sensorimotor as well as patterns of distributed lexical co-occurrences. Their model successfully accounts for a range of typical and atypical semantic cognition phenomena, including the role of executive function influences on task-specific meaning construction. However, the task-specificity is constructed through the use of homonym concepts (e.g. bank) as opposed to a wide range of concepts more generally.

In our view, this research, although very interesting, struggles to genuinely integrate so-called embodied and disembodied information due to the artificial nature of the stimuli used. We argue that the model effects observed are a direct consequence of the stimuli features artificially generated based on *a priori* assumptions of concept overlaps. We now cover a range of evidence-based semantic dimensions in the extant literature, before turning our attention to grounding concrete and abstract concepts in naturalistic scenes.

5.2.3 Dimensions of Semantic Representations

At this stage, to broaden our perspective and to lay the foundation for this chapter and the next, we opt for a highly inclusive definition of the term *semantic dimension* as a measurable extent of some meaning-related property. The three main classes of describing semantic dimensions commonly identified in the cognitive science literature are as follows: (i) *taxonomic* (e.g. animate-inanimate), (ii) *modality-specific* (e.g. vision, auditory, haptic) and (iii) *concreteness* (concrete-abstract). We find these three areas particularly relevant for grounding concrete and abstract concepts because the putative dimensions are likely to be important determinants of *how* concepts are grounded. Lastly, to critically evaluate the taxonomic perspective, we will also briefly evaluate an alternative perspective, called *lexical-retrieval*, which is not so much a perspective on describing meaning but rather a reinterpretation of neuropsychological findings, which lends support to the taxonomic dimension. In this section, we briefly outline each of these four perspectives before delving into our own computational studies of concrete and abstract concepts.

5.2.3.1 Taxonomic categorisation

Cognitive psychologists, especially those studying language, have long argued that semantic memory of concrete words is structured hierarchically (Rosch, 1975; Ebeling, 1978; Rosch & Lloyd, 1978). The first seminal neuropsychological clinical case study related to the taxonomic categorisation of semantic memory originates from Warrington and McCarthy's (1983) female patient nicknamed V.E.R who suffered a severe left hemisphere stroke. They demonstrated that V.E.R's ability to speak and follow relatively simple verbal instructions was severely impaired following the stroke, although she was able to point to pictures in response to a spoken word, revealing some sparing of semantic memory. The exciting discovery was the nature of her memory impairment. V.E.R was

significantly more successful at pointing to pictures of animate and natural objects than inanimate objects such as tools and other artefacts. This finding led them to label the condition as *category-specific access dysphasia*, which more generally refers to category-specific semantic deficits.

Subsequent evidence provided by Warrington and Shallice (1984) revealed the exact opposite pattern in four patients with halted progression of *herpes simplex encephalitis* - a viral infection of the central nervous system (Whitley et al., 1982), all of whom performed worse when recognising animate concepts than inanimate ones. Consecutive demonstrations of independent dissociations in opposite directions collectively support a double dissociation interpretation, which points to more conclusive evidence for claiming that two distinct cognitive functions might originate in different areas of the brain. However, this long-standing hallmark of a "genuine finding" in neuropsychology, dating back to Teuber (1955), has been questioned previously by Shallice's (1988) suggestion that dissociations only signify discriminations between brain regions if one first accepts the premise of *modularity* - the brain consists of distinct areas dedicated to specialised information processing. More recently, Van Orden, Pennington and Stone (2001) use the case of reading modules to show that such strict adherence to modularity is misguided, due to the inability of finding pure dissociations, which leads to the second problem of unrealistically finding purer cases, resulting in a regression towards increased fractionation and a diminishing likelihood of detecting patients with specialised damage.

However, from our perspective of understanding cognitive semantics, the double dissociation of animate and inanimate conceptual taxonomies is a particularly significant milestone given that it marks the first objectively identified unitary dimension of semantic processing, namely animacy versus inanimacy. At this stage, we move away from the dichotomous distinctions initially outlined in the neuropsychological literature (e.g. Caramazza & Shelton, 1998), because recent fMRI evidence

suggests a more psychologically plausible alternative. Sha et al. (2015) applied *representational similarity analysis* to the ventral visual cortex and found a graded activation continuum between most animate and least animate concepts. Sha and colleagues, using 480 images split across 12 categories, showed that the most critical dimension (first principal component) discriminated between inanimate and animate stimuli. Categories KEY, HAMMER and LOBSTER were at the lowest end of the animacy spectrum, while images from the categories CAT, CHIMP and HUMAN were on the opposite end of the spectrum (high animacy) and PELICAN, WARBLER, CLOWNFISH and RAY were in the middle of the first principal component. Sha and colleagues argue that the neuropsychological findings supporting a dichotomy is “the illusory result of stimulus sampling biases” (p. 665). Therefore, previous neuropsychological deficits outlining the dichotomous difference between animate and inanimate concepts might well be due to sampling from either end of the spectrum.

5.2.3.2 Modality-specific categorisation

The second framework for delineating semantic cognition also stems from the interpretations of the neuropsychological double-dissociations discovered by Warrington and colleagues (Warrington & McCarthy, 1983; Warrington & Shallice, 1984). Warrington et al. argued that different sensorimotor channels might be weighted deferentially as a function of the type of stimuli represented in semantic memory. They predicted that semantic memories of animate objects would be associated more strongly with perceptual attributes, while for semantic associations of inanimate objects, functional properties would have stronger relevance. A concrete example of this would be the comparison between the animate and inanimate objects LEOPARD and DESK, where for the animal the dominant sensorimotor channel would be visual, while in the case of the inanimate

piece of furniture, it would be the functional properties of the object (Farah & McClelland, 1991, p.341).

The first pioneering cognitive modelling experiment in the domain of semantic cognition and crucial theoretical development in the same field culminates from Farah and McClelland's (1991) landmark publication, cited nearly 1,000 times. Farah and McClelland used a connectionist architecture to model 20 random concepts (ten animate and inanimate), consisting of 24 visual and 24 verbal input units. The emphasis on randomness is two-fold. Firstly, the words represented in the hidden layer of the neural network were created arbitrarily with -1 and +1, which might appear unusual, but is logical given their modelling focus on the connectivity of stimuli-groups (living and non-living) and functional and visual semantic spaces. Secondly, one of our core arguments we have made throughout is the need for ecologically representative stimuli in cognitive modelling. Therefore, technical details like "random stimuli" are a vital point of departure for our present modelling studies. In fairness, our real-world grounding is a more manageable endeavour for contemporary cognitive modellers than was the case almost three decades ago, due to the current availability of increased computational power, the digitisation of much of our everyday lives and the ubiquity of machine learning techniques developed more recently, in particular, *deep learning*. A theoretically-motivated modelling constraint implemented by Farah and McClelland was based on the relative frequency of visual and functional properties associated with animate and inanimate concepts. Their experimental ingenuity was based on asking participants to read through dictionary definitions of living and non-living words and highlight the number of instances either functional or visual characteristics were described. This quantification was used to test Warrington and Shallice's (1984) theory of visual and functional properties respectively being more important for living and non-living objects. Indeed, the ratio for visual-to-functional for animate concepts was 7.7:1 and for inanimate objects 1.4:1,

which indicates that the visual channel was stronger for living concepts than non-living ones, even though the original assumption of functional properties being more important than visual properties for inanimate concepts was found not to be the case. These ratios obtained determined the proportion of visual and functional units in the semantic layer of the neural network, which was then subjected to gradual noise perturbation experiments to simulate brain damage by progressively lesioning the visual or functional semantic representation layer of the network from 0% (no damage) all the way to 99%. Farah and McClelland's modelling results quantitatively and qualitatively reflected the double dissociation observed in the neurological patients, where more significant damage to visual units had a more substantial impact on animate concepts, while damage to the functional units impacted the inanimate objects (Postle, 2015).

This perspective of modality-specific semantic memory is particularly important for the research presented in this chapter but also this thesis in general given its relevance to grounded cognition. A range of theories have outlined the importance of modality-specific conceptual processing (e.g. Barsalou et al., 2003; Niedenthal, 2007; Van Dantzig et al., 2008; Gallese & Cuccio, 2018), and the connectionist PDP approach implemented by Farah and McClelland (1991) highlights the theoretical importance of cognitive modelling to mechanistically theorise about semantic cognition.

5.2.3.3 Lexical-retrieval categorisation

The third framework for explaining semantic processing is called *lexical retrieval*, advocated by Damasio and colleagues (1996), and subsequently supported by Grabowski, Damasio and Damasio (1998). At its core, Damasio et al. (1996) use neurological and neuroimaging evidence to claim that the *lateral temporal-lobe* is critical for the retrieval of words, and does so in a taxonomically structured manner. This argument is in stark contrast to the previously discussed interpretations of the double-

dissociation of modality-specific accounts of animate and inanimate semantic dementia given the emphasis is on the retrieval process from a mental lexicon. Postle (2015) argues that this might be related to Damasio's neurological background, which places emphasis on observable symptoms and is more in favour of characterising symptoms as primary progressive aphasia as opposed to semantic dementia, which is more common in the psychological literature. Damasio et al. (2004) further support their earlier claims using lesion and neuroimaging studies, which shows that the left temporal lobe contains partially segregated higher-order cortical regions associated with the retrieval of concrete words from different conceptual categories. Tranel (2006) further extended this by observing that damage to the *left temporal polar* (TP) cortex is associated with the impaired naming of unique landmarks, in support of the mediatory role of the TP in retrieving specific classes of stimuli.

Until now, our overview of the taxonomic and modality-specific perspectives has focused on dichotomous or unidimensional properties of semantic representations. The lexical-retrieval interpretation was briefly outlined as an alternative interpretation of the double dissociation findings from Warrington and Shallice (1984). Next, we turn our attention to the concreteness continuum, an evolution from previously held dichotomous standpoints.

5.2.3.4 The Concreteness Continuum

Yarmey and Thomas (1966) showed that concrete words such as DOG or TABLE are different to abstract words like LOVE. In their experiment, where participants had to learn two lists of abstract and concrete nouns, they found that learning concrete nouns was easier than learning abstract nouns by comparing the recall accuracy. This study marked the beginning of the *concreteness effect*. This advantage of concrete words over abstract words is found across a diverse range of cognitive domains. Gilhooly and Logie (1980) conducted a large-scale study of 1,944

words and metrics such as *age-of-acquisition* (AoA), *familiarity* (FAM) and *concreteness* (CNC) and created the first comprehensive dataset of all these measures. Gilhooly and Logie found a strong positive correlation of 0.93 between AoA and CNC.

Warrington (1975) was the first to consider a reversal of the concreteness effect. One of the patients examined was AB, formerly a high-level civil servant who performed very well on a verbal IQ test (122) and was not showing any signs of dysphasia despite using a restricted vocabulary. In the first set of neurological tests evaluating picture and word recognition, AB had significant difficulties with the visual version of the task compared to the auditory one. AB was able to define an object name but struggled to describe the visual features of the object. Warrington noted that for patients AB, and EM and CR, this so-called *object-agnosia* was characterised by two forms of response errors when asked to describe the visual properties of objects. The first type of error was distinguished by a tendency to define concepts by reference to superordinate level categories. For example, Warrington notes, that the word HAMMER would be described as “some kind of tool” (p. 642). The second type of error was characteristic of defining the original concepts (e.g. DONKEY) by reference to another object within the same taxonomy, such as a HORSE. Even in subsequent visual recognition by forced choice tests, where there was simply a yes/no answer to questions such as “is this a bird?”, AB performed very poorly. Strikingly, however, AB’s word recognition performance, as measured by open-ended verbal responses, for abstract words was normal. AB did not have difficulty defining words such as PACT (AB’s response: “friendly agreement”) or TAME (AB’s response: “an animal not behaving wildly”), whereas for concrete words like HAY, POSTER and NEEDLE, typical responses were “I’ve forgotten” or “no idea”. Warrington and Shallice (1984) further supported this reversed concreteness effect, with patient SBY, whose semantic deficits closely matched that of AB’s.

Breedin, Saffran and Coslett (1994), demonstrated another neuropsychological case where there was a reversal of the concreteness effect. Their patient DM, who was suffering from semantic dementia due to neural atrophy in the left temporal lobe, had profound difficulties with object names, even though much of his knowledge of abstract nouns and verbs remained intact. This dissociation led Breedin et al. to argue against a “quantitative interpretation” of the concreteness effect, which they stated as the superiority of concrete words over abstract words in conceptual processing (p.617). More recently, Shallice and Cooper (2013) discuss double dissociation as well as neuroimaging studies in support of partially separable abstract and concrete conceptual representation systems.

Evidence on word concreteness effects has amassed significantly in recent years in a range of domains such as reading comprehension, word recognition and judgements (e.g. Walker & Hulme, 1999; Allen & Hulme, 2006; Witherby & Tauber, 2017). Nonetheless, only two mechanistic candidate theories are available, which are Schwanenflugel and Shoben's (1983) *Context Availability Model* (CAM) and Paivio's (1991) *Dual Coding Theory* (DCT). A central tenet of CAM is that due to the easier accessibility and the increased availability and richness of contextual information, concrete words can be dissociated from abstract words, which also accounts for the advantage of concrete words given their conceptual wealth. Schwanenflugel and Shoben's theory is based on lexical decision time studies, so their use of the term *context* refers specifically to the linguistic context, which is different from our broader use of the term. They presented abstract and concrete sentences, either with or without a context paragraph; in the condition without context, participants took longer to read abstract sentences than concrete ones, while in the context paragraph condition, participants took equally long to read the abstract and concrete sentences.

The second theory, Paivio' DCT (1971, 1991), is based on two parallel semantic systems, one for *verbal semantics* and the other for *imaginal semantics*. In DCT, the advantage of concrete words is accounted for by co-

activation of both the verbal and imaginal semantic systems, while in the case of abstract words, only the verbal semantics system is activated. A commonality shared by both CAM and DCT appears to be in their same predictions for increased processing efficiency or speed for concrete words compared with abstract words, be that reflected in faster reaction times or lower error rates for the former. However, these two theories, especially in the case of DCT with its parallel semantic systems, are also premised on the idea that there seems to be a sharp dichotomy between concrete and abstract words, and that a continuum between the two is lacking. This will be particularly relevant for our re-analysis in the next section, as it questions the validity of this premise.

5.2.4 Concreteness Revisited

Brysbaert, Warriner and Kuperman (2014) conducted a large-scale study with over 4,000 participants, to collect concreteness ratings on a 5-point Likert scale across 40,000 common English word lemmas, and published their dataset. We re-analysed this data to produce a simple histogram of concreteness ratings, where the concreteness ratings are split into 100 distinct bins (see *figure 5.2A*). The concreteness ratings are distributed *continuously* across the Likert-scale response variable even though there is evidence to support a weak bimodal distribution, with the lower end of the spectrum being more leptokurtic than the higher end. This helps our interpretation of a concreteness continuum. However, we find, that even stronger and more convincing evidence emerges from our second chart (see *figure 5.2B*), where we plot the standard deviations of the ratings along the concreteness spectrum. Our analysis shows that words on either end of the concreteness spectrum have lower variations in their scores, presumably based on their consistent interpretations. Some of the words with the smallest standard deviations of zero are TULIP, BIRD, BIKE, FISH, HANDSAW and even YO-YO among a list of 280-word lemmas, which also have the highest concreteness rating of five.

Similarly, for words with the lowest concreteness ratings between 1.04 and 1.20, the standard deviation was typically less than 0.4, including words such as SPIRITUALITY, ENLIGHTENING, and IDEALIZE. Words in the middle of the spectrum show the highest level of variation, which gradually increases from both concrete and abstract ends. This strongly indicates that word concreteness is continuous. Based on these descriptive statistics, it would appear that the concreteness dichotomy can be primarily attributed to experimenters selecting words from either extreme, which are more consistent in their meanings, neglecting the middle of the spectrum, leading to results focusing either on abstract or concrete words. This might also explain why some of the earlier studies (e.g. Troche et al., 2014) found prominent bimodal distributions in concreteness ratings, while a continuity emerges in more recent and large-scale studies spanning a wider range of words (e.g. Brysbaert et al., 2014).

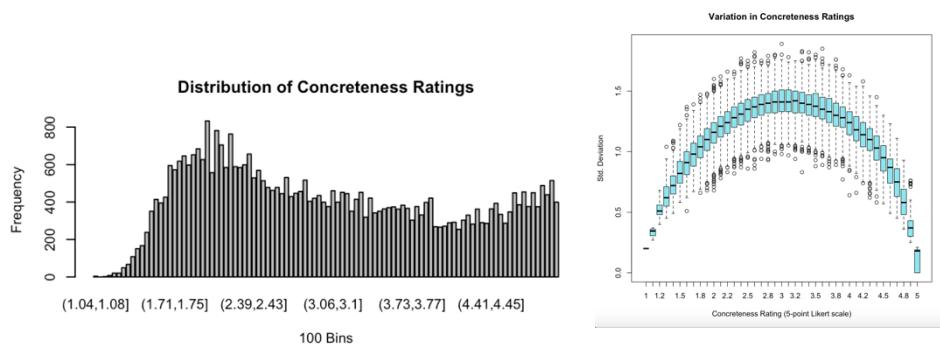


Figure 5.2: (A) Histogram of the concreteness ratings of 40,000 word lemmas (B) Boxplot of the standard deviations of the same concreteness ratings. See *Appendix D* for enlarged images (p. 334).

5.2.5 Role of Emotions in Grounding Abstract Concepts

Emotions are a central and critical aspect of cognition, despite a long history of cognitive sciences ignoring the role of affective processing as a key property of cognition. More recently, this has also led to computational cognitive theorists extending their more traditional models with *emotion modules* across a range of cognitive architecture like ACT-R (Laird, 2008) and SOAR (Dancy, 2013). Developmental theorists such as Bloom (1998)

have also shown that even though some researchers have long investigated the use of emotion words in children's development (e.g. Morton & Trehub, 2001), emotional development is actually a precursor to the development of language, and that the use of emotion words only occurs at approximately two years of age. However, the rate of acquisition of emotional words steadily increases between two and three years of age (Wellman et al., 1995) and might provide a bootstrapping mechanism for the acquisition of abstract words (Kousta et al., 2011). Fay and Maner (2012) provide an account in which environmental factors in early child development can contribute to early non-verbal categorisation. They note that Bowlby (1969) suggests the importance of physical proximity between infants and caregivers and the association with warmth, and how this early spatial proximity can later in development be used for conceptualising more abstract concepts like *closeness* and *intimacy* (p.1369). Other research has extended such spatial proximities to other abstract concepts like *warm friendship*, through metaphoric extensions (Williams & Bargh, 2008), though such findings are beyond our focus on emotions and their contributory role in grounding abstract concepts.

Hedonic valence, although related to emotions, more specifically refers to the binary opposition of good versus bad, and their corresponding strength has been widely cited as a potentially important criterion for, at least partially, constituting the meaning of abstract words. Altarriba, Bauer, and Benvenuto (1999) initially suggested the importance of valence interacting with concept concreteness, although their suggestion was to create an additional distinction of emotional words in addition to concrete and abstract concepts. Kousta and colleagues (2011) further developed the descriptive postulates of Altarriba et al. (1999). Kousta et al. hypothesise that the two dominant cognitive theories, *dual coding theory* and *context availability model*, are both inadequate for explaining these results due to their lack of incorporating experiential information in grounding abstract concepts. Indeed, their study supported their hypothesis and marked an

important discovery for not only cognitive semantics but also *grounded cognition*.

Building on previous work (Kousta, Vinson, & Vigliocco, 2009) demonstrating the processing advantage of concepts with emotional associations, Kousta et al. (2011) ran a series of experiments along with predictive modelling of a large-scale database (*English Lexicon Project*; Balota et al., 2007). The dual coding theory and the context availability model are respectively based on the empirical findings of *imageability* and *context availability*. Imageability refers to the ease with which a given concept can elicit mental images of things or events, and is commonly measured on a 7-point Likert scale ranging from *low* (1) to *medium* (4) to *high* (7). For example, concrete words such as *apple* have significantly higher imageability ratings than abstract words like *justice*. Unsurprisingly, concept concreteness and imageability are highly correlated (Toglia & Battig, 1978).

Similarly, context availability refers to the ease with which a given context or situation can be elicited for a given concept (Kieras, 1978). Kousta et al. (2011) revealed that once both imageability and context availability were controlled, abstract word processing has a small but significant advantage. Moreover, Kousta and colleagues also demonstrate that this representation difference can be explained by variations in valence ratings, which provides a new unidimensional view of the cognitive semantic system, based on concept concreteness. Kousta et al.'s finding suggests a novel hypothesis, which they name *embodied abstract semantics*, while simultaneously discrediting *dual coding theory* and the *context availability model*, both well-established theories of accounting for semantic and processing differences between concrete and abstract concepts.

Vigliocco et al. (2013) extend the behavioural evidence of Kousta et al. (2011), using a neuroimaging study which provides evidence for activations in the *rostral anterior cingulate cortex*, known for its role in *emotional conflict resolution* (Etkin et al., 2006) through the modulation of

hedonic valence. Additionally, correlation analysis on 1,400 words also revealed that abstract words are associated with higher emotional ratings. In summary, the above findings across a range of different studies point towards a strong possibility of emotions being especially crucial for grounding abstract concepts, although there is currently no computational or mechanistic account of this grounded phenomenon.

5.3 Computational Study I: Evaluating PSVs using BrainBench

5.3.1 Objective

Xu, Murphy and Fyshe (2016) developed a small test suite for evaluating distributed semantic machine learning models, based on the interrelations of brain activations from Mitchell et al. (2008) and Sudre et al. (2012). Xu et al. found that words that are correlated behaviourally (e.g. reaction times) are also correlated neurally (i.e. have similar neural activations). Their *BrainBench* measure computes a correlation between a given distributed model and their benchmark neuroimaging dataset on the same set of concepts. Although all 60 concepts in the test set are comprised of concrete words across a range of categories from *body parts* and *vehicles* to *vegetables* and *tools*, we use this as a benchmarking exercise for evaluating *Perceptual Scene Vectors* (PSVs) against commonly used distributed linguistic models. These 60 words are then used throughout the studies in this chapter, although more abstract words are also added for studies II and III. Similar to distributed linguistic models, we predict a positive correlation between neuroimaging data and our PSV model.

5.3.2 Methodology

Using Xu et al.'s (2016) web-based interface⁸, we individually uploaded a range of off-the-shelf linguistic representations, some of which Xu et al. provide on the website (Skip-gram, Glove, RNN, Global and Cross-lingual), while also including LSA models trained separately on the *TASA corpus* and *Wikipedia*. All of these distributed models included the same range of dimensions as Xu et al. For our additional LSA models, we included LSA 50, 100, 150, 200, 250 and the standard 300 dimensions. Our PSV representations were generated using our *semi-automated* methodology same as implemented in chapter 4, with only two differences. Firstly, we increased the *number of images* from 10 to 20 per concept from which PSVs were extracted and secondly, we increased the hidden layer from 20 to 100 neurons. We manually *excluded* image search results that we qualitatively determined did not satisfy the criteria of an *ecologically valid* scene one is likely to encounter in everyday life (e.g. brand logos or artwork). For example, for the word *eye*, we excluded anatomical diagrams, while for *legs* we discarded an image of a fitness device, which did not include legs.

⁸ http://www.langlearnlab.cs.uvic.ca/brainbench/#filter_vector

5.3.3 Results

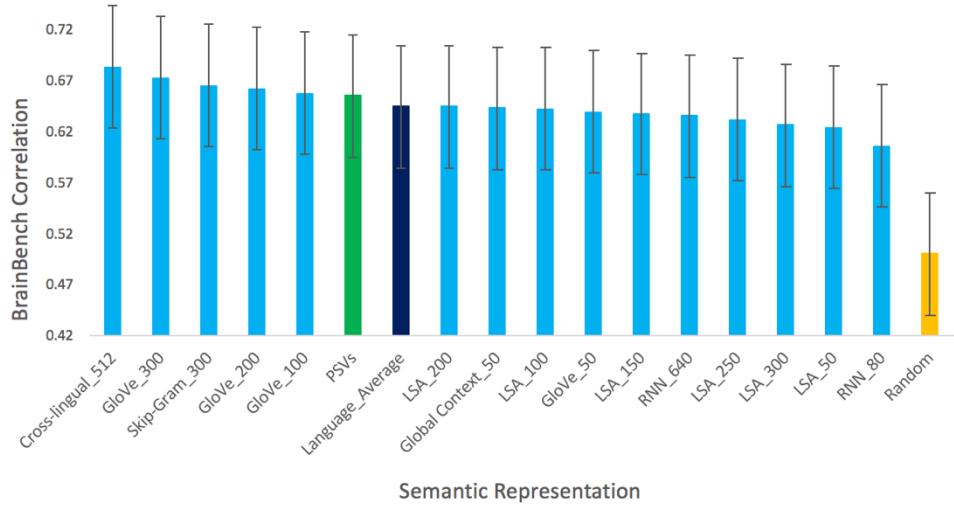


Figure 5.3: A bar chart of the BrainBench results across a range of distributed “off-the-shelf” representations and our Perceptual Scene Vector (PSV). The semantic representations are ordered from highest-to-lowest correlations.

Although Xu et al. (2016) provide both fMRI and PET BrainBench correlation coefficients, we did not find any interpretable differences in the two highly correlated scores ($r = 0.84$, $p < 0.01$). Therefore, we take the average of the two correlations and report this in *figure 5.3*. According to the BrainBench comparisons, both linguistic and our PSV representations perform comparably. Our tests with random vector representations also confirm Xu et al.’s random benchmark of $r = 0.50$, but we further establish that at a 95% confidence interval⁹, the variation in r is between ± 0.06 , which prevents us from interpreting differences in the variations of the scores between the language-based and scene-based distributed representations. The best distributed linguistic representation is Cross-lingual ($r = 0.68$), while the worst performing linguistic representation is RNN with 80 dimensions ($r = 0.61$). All language-based and PSV representations (averaged across ten trials of neural network training with different initialisations) perform better than Xu et al.’s random benchmark

⁹ We measured r for the *random* vector 100 times, and 95 times it is between ± 0.06 .

of $r = 0.50$. These results show that although from this benchmarking exercise we cannot draw any conclusions on whether or not PSVs perform better or worse than language-based distributed models, our visually grounded representations perform equivalently.

5.4 Computational Study II: Grounding Concrete to Abstract Concepts

5.4.1 Objective

In this computational experiment, we aim to investigate the generalisability of our PSVs across the concreteness spectrum, in order to understand the strengths and limitations of PSVs across a broader range of concepts on the concreteness spectrum. We use concepts across the spectrum and split them into *concrete*, *intermediate* and *abstract* concepts. Given that PSVs are semi-automatically extracted object-level characteristics from naturalistic scenes, we expect the quality of semantic representations to gradually increase along the concreteness spectrum (least to most concrete). In this second study, the *quality* of representations refers to a qualitative interpretation of the category structure of the correlation matrix resulting from visualising the hidden layer neurons from the trained neural network.

5.4.2 Methodology

We extend the 60 concrete concepts from the first study, with addition of intermediate ($n = 20$) and abstract concepts ($n = 20$) from Brysbaert et al. (2014). The concrete concepts have a mean concreteness rating of 4.87, with a standard deviation of 0.12. Intermediate concepts have a mean concreteness rating of 3.10, with a standard deviation of 0.41, while abstract words have a mean rating of 2.20, and a standard deviation of 0.39. Both intermediate and abstract concepts are selected by first choosing five

distinct words, and then supplementing these with their three close neighbours using the TASA-trained LSA space. This is done to ensure that our intermediate and abstract words could be compared meaningfully with the LSA space in study III. Therefore, we aim at avoiding the selection of arbitrary concepts, which would make it more challenging to compare strong associative relationships. We reason that randomly selecting abstract words is likely to lead to a set of words with low comparability and therefore problematic evaluations when comparing interrelations. In study III, we also evaluate the concept representations on a one-to-one and one-to-category basis, for which we need a set of coherently and meaningfully grouped concepts. The same method as in study I is used to generate PSVs.

5.4.3 Results

The visualisation of the hidden layer units (see *figure 5.4*) of the concepts generated using PSVs shows strong category-level differentiation across concrete concepts and even some hierarchical groupings as in the case of *body parts* and *clothing*. This is also the case for the first two groups (closest to concrete concepts) of concepts with intermediate levels of concreteness ratings, which are also clearly differentiated. However, from the third intermediate group onwards (*maths*, *physics*, *intelligence*, and *numbers*), concepts become less coherently clustered within their specific groups, while simultaneously becoming more strongly correlated with other intermediate and abstract concepts. In order to qualitatively understand the possible determinants of this, we investigate the extracted features embedded in the PSVs.

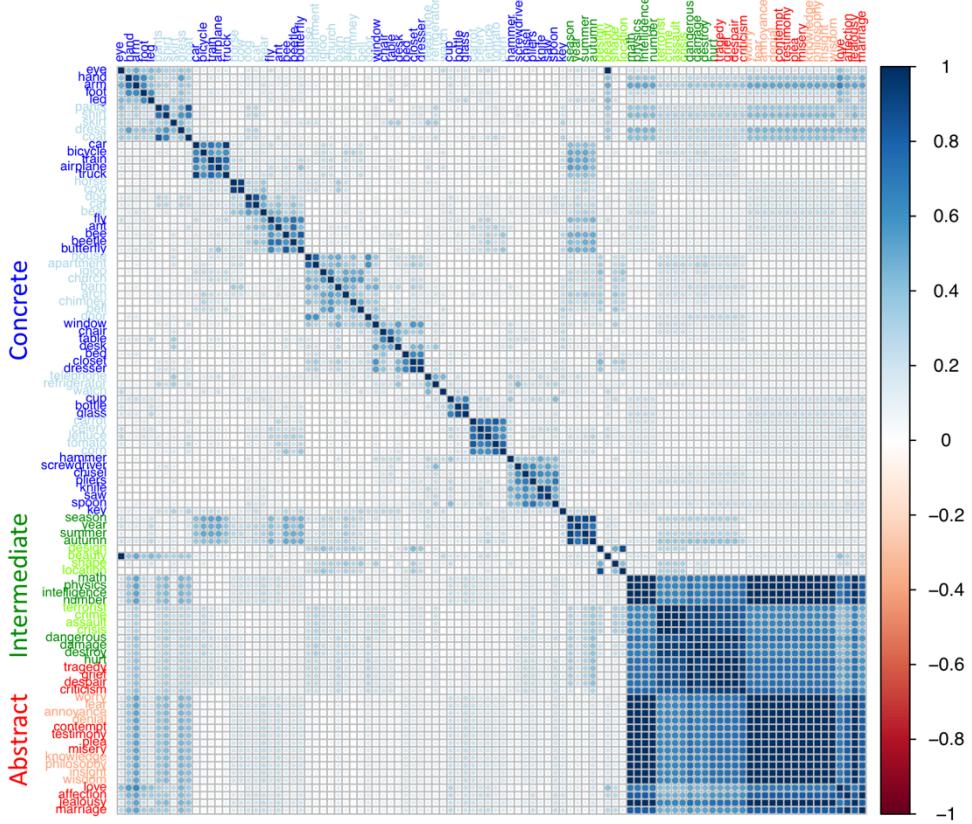


Figure 5.4: A correlation plot of the PSV’s hidden layer representations. Concepts are grouped into concrete (blue), intermediate (green), and abstract (red) groupings. Within both the intermediate and abstract groups, concepts are grouped into the LSA-based “concept clusters”, and these are highlighted by respectively alternating between darker and lighter shades of green and red. Similarly, we also alternate darker and lighter shades of blue for concrete words but use the original order of concepts used by Xu et al. (2016). See Appendix D for enlarged images (p. 335 - 336).

The PSV features reveal that intermediate and abstract concept features have high levels of global similarities. This includes features such as *indoor*, *person* and objects commonly found in indoor environments, such as *chairs*, *tables* and *computers*. The correlation plot also reveals that PSVs are poor at discriminating between more nuanced intermediate and abstract groups of concepts like {*terrorist*, *crime*, *assault*, *crisis*} and {*dangerous*, *damage*, *destroy*, *hurt*}. This is more apparent for larger groups of concepts in the abstract set, where *knowledge* is no more similar to *wisdom* than to *contempt*. These results show that although PSVs are sufficiently capable of representing concrete semantic concepts, the same level of grounding in real-world naturalistic scenes is insufficient for grounding some intermediate and most abstract words tested in this experiment.

5.5 Computational Study III: Extending PSVs with Emotions

5.5.1 Objective

In our final study, we extend our PSVs with an additional 8-bit vector representation of emotions grounded in the natural scene photographs, which we name *scene2vec*. As discussed earlier, there has been an increasing focus on the role of emotions in grounding abstract concepts in cognitive science. However, we aim to explore this from both a mechanistic computational and a grounded cognition perspective. Our research objective is to investigate the computational feasibility of better representing abstract concepts using semi-automatically extracted emotional information from real-world photographs. In this study, we can evaluate narrower hypotheses as a result of quantifying the quality of semantic representations. We operationalise this via cross-comparisons of the correlation matrix obtained from the hidden layer neurons of the network trained on PSVs (study II) and PSVs + emotions (study III) with a linguistic “ground truth” correlation matrix (LSA space). However, we have two accuracy metrics. The first accuracy metric measures *concept-to-concept* level matches, for example, concept *eye* from PSVs being most strongly correlated with the *eye* from the LSA correlation matrix (details in next section). The second metric consists of correlations between individual concepts like *hand* from PSVs being most strongly correlated with the LSA factor *body parts*, comprised of {*eye*, *hand*, *arm*, *foot*, *leg*}. This second metric is a more general *category membership* measure, while the first is a narrower *concept-specific* one.

Despite the exploratory nature of our overall study, in this experiment, we have a range of directional hypotheses. These not only narrowly focus on the specifics of *scene2vec*, but rather, compare *scene2vec* and PSV representations on both concept- and category-level accuracies,

split across concrete, intermediate and abstract concepts. Below we outline our five core hypotheses for this, along with supporting evidence and reasoning for each.

- I. For both PSV and scene2vec representations, category-level accuracies will be higher than concept-level accuracies.
 - o There is a decreased probability of misclassification (21 categories versus 100 concepts) for category-level matching, which will provide a “lower boundary” accuracy threshold.
 - o PSVs encode the statistical regularities of objects within natural settings, and scene2vec additionally contains emotional information. We, therefore, predict that there should be sufficient meaningful overlap between these regularities across a category (but not within concepts) irrespective of where the concepts sit on the concreteness spectrum.
 - o Algorithmically, we reason that it will be easier to link grounded concepts to more generic linguistics factors (derived from LSA) as opposed to the concept-specific associations derived from the LSA space.
- II. PSV representations will have the highest level of concept and category accuracies for concrete concepts, followed by intermediate, and then abstract concepts.
- III. Previous research (e.g. Grondin, Lupker, & McRae, 2009) has shown that concrete concepts are more strongly associated with objects in the environment compared to abstract concepts. We reason that intermediate concepts are likely to fall between these two extremes.
- IV. Scene2vec representations will have equally high levels of concept and category accuracies across abstract, intermediate, and concrete concepts.
 - o Borghi et al. (2017) have empirically shown that abstract concepts are more strongly linked to introspective associations and emotions. Therefore, in our scene2vec representation, concrete concepts will be successfully grounded in object co-occurrence information while abstract concepts will be grounded in object co-occurrences and emotional data.

V. Across concrete, intermediate and abstract concepts scene2vec will have a higher concept- and category-level accuracies compared to PSVs.

- Given that scene2vec representations contain both object co-occurrence as well as emotional information, we postulate that these different formats of cognitive representations will be mutually reinforcing across the concreteness spectrum.

5.5.2 Methodology

We use the same stimuli as in experiment II, and the LSA space used is also identical. The additional manipulation in this experiment consists of the extra 8-bits of emotional information. We use Microsoft's Cognitive Services' Emotion API, a cloud-based, research-grade emotion recognition software solution, optimised for real-world scalable emotional classification tasks. We collected eight emotional measures, all ranging from 0 to 1 (highest emotional rating), see *figure 5.5*.

The scene2vec representations' *Perceptual Scene Vector* (PSV) component is generated using the same methodology as used in chapter 4 and is identical to experiment II of this chapter. We once again use photographs representing an ecologically valid cognitive data representation of the visual world and run these through Zhao et al.'s (2017) pre-trained pyramid scene parsing network (PSPNet).

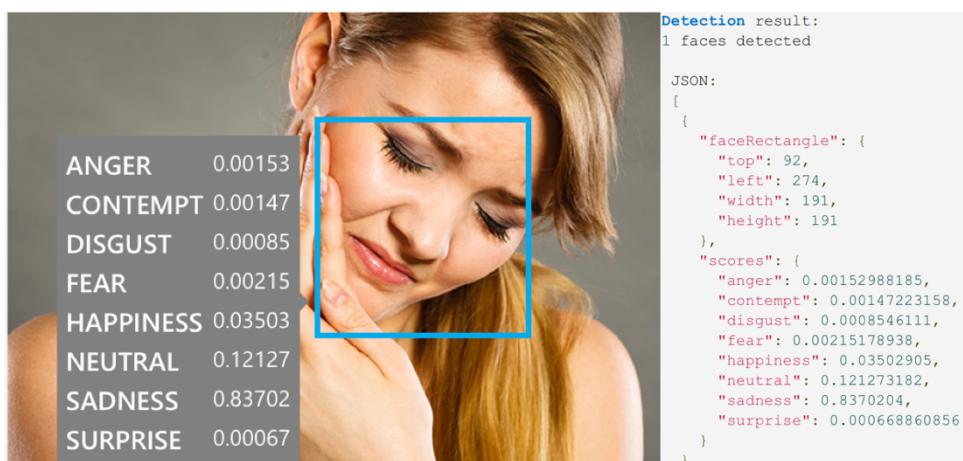


Figure 5.5: Example output of the emotion detection and the resultant JSON file. Emotional scores are generated individually across multiple faces per image.

We have encountered two specific challenges when creating scene2vec representations. One of them is the issue we also previously discussed in chapter 4 for our PSVs, where we filter out “animated” or “commercial” logos that appeared in the automated Google image downloads. However, a secondary issue we have encountered in this experiment is that many photos do not have visible faces, which impacts the quality of the emotions extracted. Therefore, we constrain the image stimuli so that at least 4 out of 20 of the photos have visible faces *if and only if* the original 20 randomly selected photos have a person present but without their face being visible (e.g. facing sideways).

In *figure 5.6* we illustrate our approach to quantifying similarities between our grounded and Latent Semantic Analysis (LSA) representations. This method is based on Kriegeskorte et al.’s (2008) Representational Similarity Analysis (RSA), which quantifies the similarities between neuroimaging, behavioural and computationally-derived representations. Kriegeskorte et al.’s technique computes *representational dissimilarity matrices* (RDMs), using correlation distances (1 - correlation) within a given representational format (i.e. using concepts’ neural signatures). This analysis quantifies first-order isomorphisms between concepts. Critically, Kriegeskorte and colleagues compute second-order isomorphisms by comparing the dissimilarity matrices of neuroimaging-based and a theoretical, computational model. This is statistically operationalised by calculating a correlation coefficient between the two RDMs, which quantifies the correspondence between the neuroimaging and computational representations.

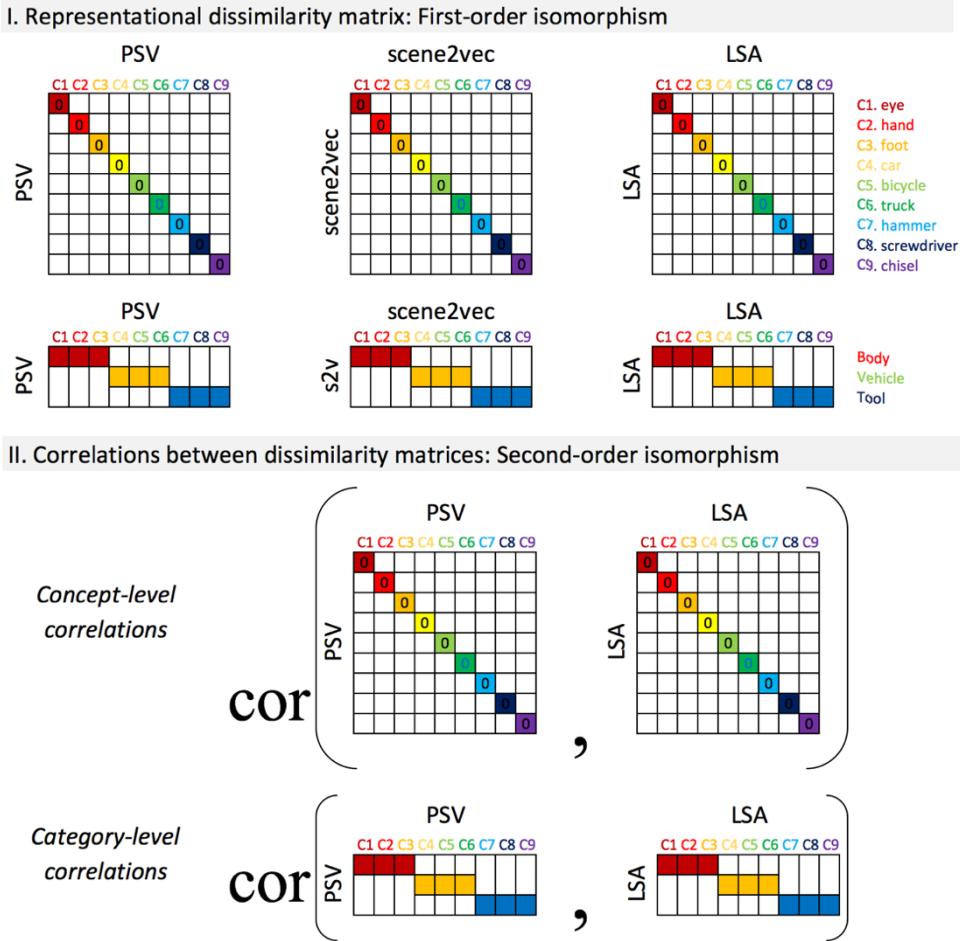


Figure 5.6: Schematic overview of PSVs, scene2vec and investigating the correspondence with LSA by correlating their respective representational dissimilarity matrices (RDMs). Concept-level correspondences are a more conservative measure and refer to a direct concept-to-concept mapping. Category-level correspondences, on the other hand, less restrictively, consider a match to be a concept-to-category association.

In our study, we match our grounded semantic representations (PSVs and scene2vec) with the well-established linguistic benchmark of LSA. Like Kriegeskorte et al., we investigate second-order isomorphisms, but in our case, between two computational representations, see *figure 5.7*. We compute RDMs using the hidden layer representations of PSVs, scene2vec and LSA's 300 dimensions across all concepts. *Concept-level accuracies* are determined using an LSA-based RDM for all concepts. A secondary measure of accuracy is our *category-level metric*, which is also based on an LSA RDM, but with 21 *a priori* LSA factors, each corresponding to categories like *body*, *vehicle* or *tool*.

Both concept- and category-level accuracies are split between concrete, intermediate and abstract concepts, resulting in 12 comparisons (see *figure 5.7*). These two approaches to measuring accuracy are based on our assumption that although grounded representations might be suitable for more concrete concepts, they will be less appropriate for more abstract ones. However, we also predict that our scene2vec representation will be superior at discriminating concepts at a higher-order categorical level, while not being as valuable for more fine-grained differentiation of concepts.

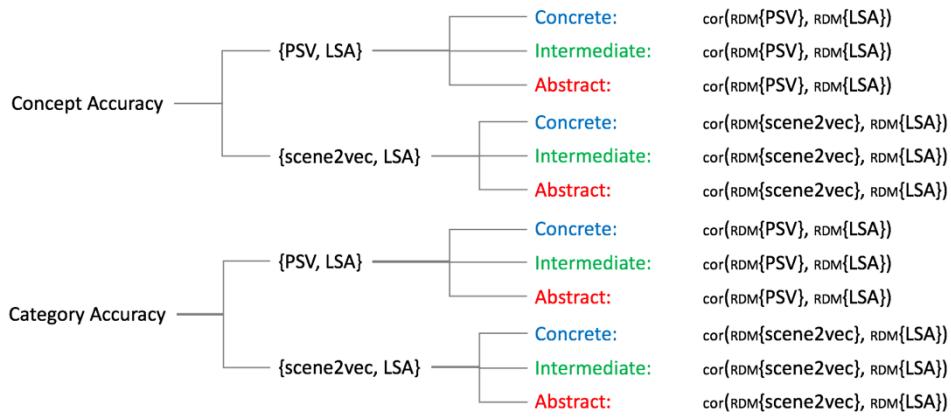


Figure 5.7: Outline of the 12 different comparisons between grounded and language-based representations, using our adapted *representational similarity analysis* (RSA). Correlations (cor) between the different representational dissimilarity matrices (RDMs) lead to concept- and category-level correspondences.

5.5.3 Results

A qualitative visual inspection of the correlation plot (corrplot) in *figure 5.8* shows that for concrete concepts the interrelations seem very similar to those based on PSV (*figure 5.4*), yet for both intermediate and abstract concepts, there are more meaningful associative relations. The most noticeable difference is that for scene2vec's corrplot, the intermediate and abstract concepts do not form as distinctive a superordinate cluster as is the case with PSVs. This suggests that more nuanced associative intermediate and abstract concept relations are captured by our addition of emotional information to PSVs (creating scene2vec). Interestingly, although

slightly weaker compared to PSVs, intermediate and abstract concepts remain strongly correlated with the concrete concepts in the *body parts* and *clothing* groups. This is understandable given that classifications of *people* and similar *indoor* scenes are visually present for most of these concepts. However, a dominant finding across both PSV and scene2vec representations is that two distinct groups are generated for more concrete and more abstract concepts respectively, with intermediate concepts being more similar, on average, to abstract concepts.

The hierarchical cluster analysis, depicted in *figure 5.9* further supports our interpretation that scene2vec and PSV representations are largely identical for concrete concepts, although there are more associatively meaningful relationships for intermediate and abstract concepts in scene2vec. Both dendograms have similar absolute distances across all the concepts, but in the case of PSV representations, for abstract and some intermediate concepts there is very little differentiation as is shown by several concepts with a difference of zero in their respective Euclidean distances. Contrasting this to the tree structure of the scene2vec representation, we can see that although there are still more abstract and intermediate concepts with zero differentiation, these cases are not as prevalent. Moreover, the level of differentiation for abstract and intermediate concepts, as reflected by the more nuanced hierarchical relationships in the scene2vec representation, suggests that the inclusion of emotional information does predominantly support the grounding of more abstract concepts.

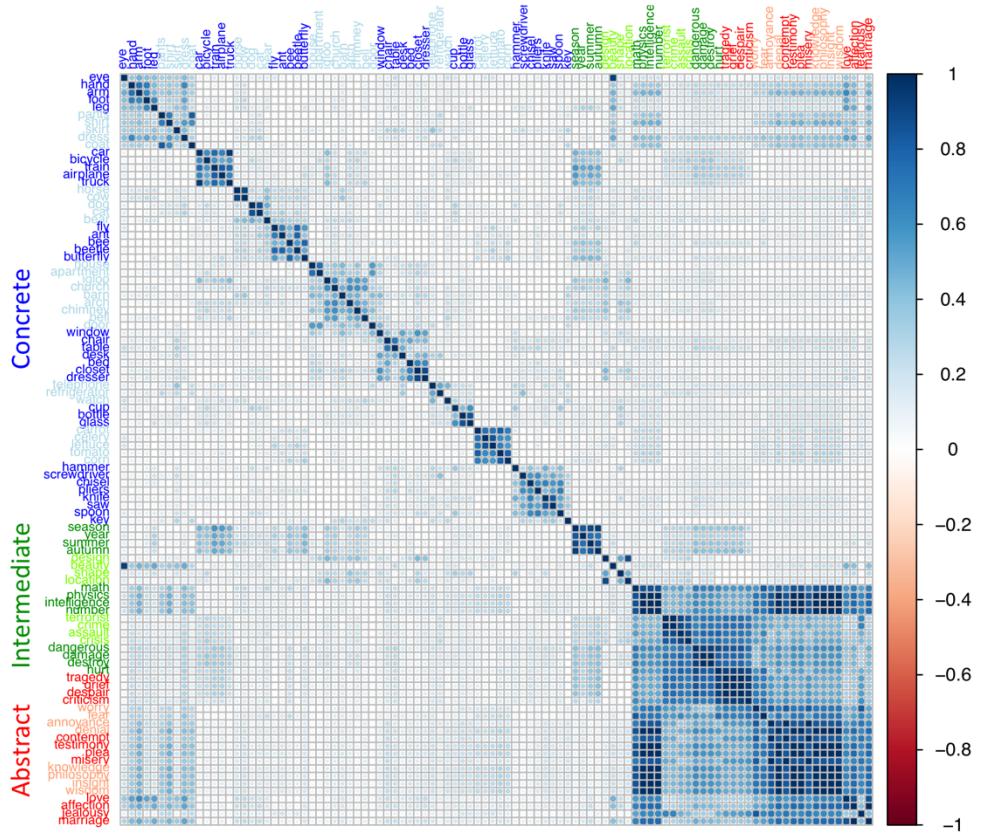


Figure 5.8: A correlation plot of scene2vec’s hidden layer representations. Concepts are once more grouped into concrete (blue), intermediate (green), and abstract (red) groupings. See *Appendix D* for enlarged images (p. 337 - 338).

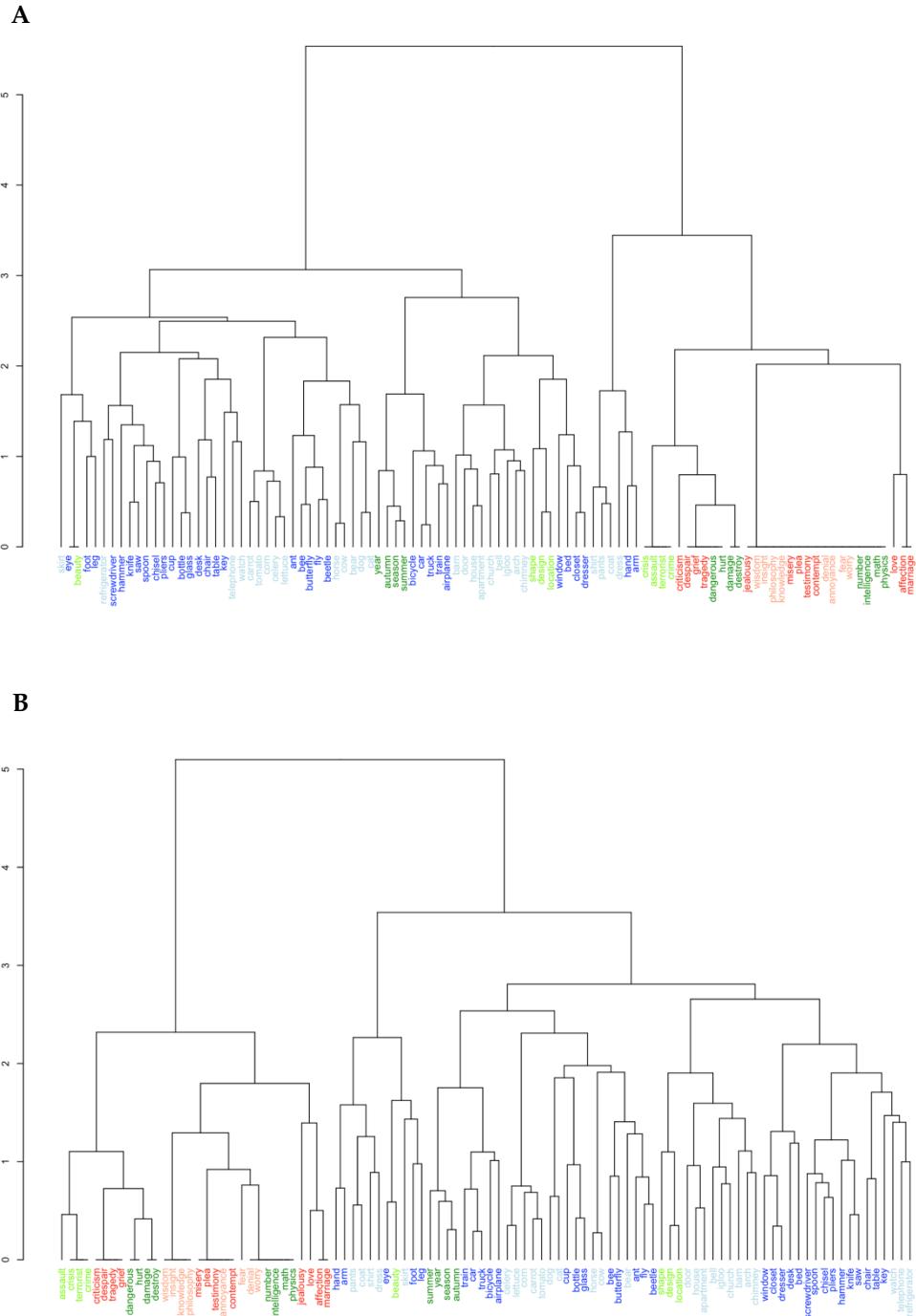


Figure 5.9: Hierarchical cluster plot of the hidden layer neurons representing the semantic associations of PSVs (A) and scene2vec (B) representations. Concepts are once more grouped into concrete (blue), intermediate (green), and abstract (red) groupings. See *Appendix D* for enlarged images (p. 339 - 340).

Our final results compare PSV and scene2vec representations with those of the distributed linguistic associations from the LSA 300-dimension model (see *figure 5.10*). The main finding is that across all grounded

conditions, concrete words have a high degree of correspondence with LSA. Therefore, concrete concepts' representations grounded in naturalistic images, irrespective of whether PSV or scene2vec distributed representations are used, is similar to those from LSA. However, for scene2vec concrete representations, the accuracy for both concept- and category-level correspondences is approaching the maximum correlation, while it is slightly lower for the PSV scenarios, indicating a slight but consistent advantage of scene2vec even for concrete concepts. Relatedly, the intermediate and abstract concept correlations are considerably lower across all conditions, although this is particularly the case for the PSV representations, indicating that PSVs are very poor distributed representations for capturing the semantic variations of more abstract concepts. In the case of PSVs, there is a gradual decrease in both concept and category correlations from more concrete to abstract concepts. This trend is not found for scene2vec representations, where intermediate and abstract concepts both have similar levels of correlation with LSA, although considerably higher correlations at the category level.

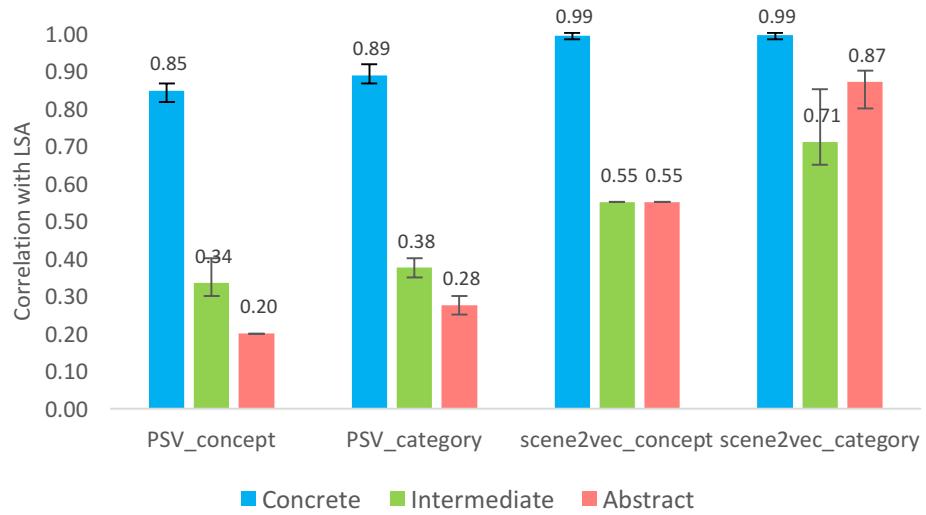


Figure 5.10: Concept- and category-level correlations of PSV and scene2vec representations with LSA-300. Bar chart colours depict concepts grouped within the three concreteness categories. The error bars show the range of correlations obtained across ten trials conducted for the matching tests, where the PSV/scene2vec neural network is initialised with different random seeds.

Our scene2vec representations capture more LSA-like semantic regularities in meaning than PSVs do, especially at the category-level, where the lowest correlation is 0.71. A somewhat unexpected result is the similarity in both concept and category accuracies for PSV representations across all three concreteness categories. Collectively, the differences in the accuracies in our results provide support for scene2vec representations, which are superior at capturing semantic associations, with their additional emotion information, comparable to language-based distributed associations, although this performance declines for intermediate and abstract concepts. The scene2vec representations also have higher correspondences at the category-level than at the concept-level. These results collectively suggest a range of strengths but also strong shortcomings in our scene-based grounding of semantics of more abstract concepts, which we discuss next.

5.6 Discussion

Our present study is the first exploratory investigation of mechanistically evaluating the scene-based grounding of concepts. Moreover, we do this across the concreteness spectrum, with a particular focus on representations with (scene2vec) and without (PSV) emotional information. All the information encoded in the input representations of the studies is extracted either automatically or semi-automatically (some irrelevant photos are excluded) from naturalistic photographs. We also present a small portion of re-analysed data from Brysbaert et al. (2014) to further support the notion of a concreteness continuum as opposed to a dichotomy, which others in cognitive modelling and robotics have also adopted (e.g. Cangelosi & Stramandinoli, 2018). The research objective of the present chapter can be broadly classified as an exploratory investigation, with the hope of helping to formulate more specialised predictions for future studies. Nonetheless, we do have four overarching

hypotheses that we have investigated across our computational experiments.

Our first prediction is that category-level accuracies will be higher than concept-level accuracies for both PSV and scene2vec representations. This first hypothesis is not fully supported by our results because despite that the results are in line with our predictions for scene2vec representations, it is not the case for PSVs. This does, however, suggest that the additional emotional information is more helpful at discriminating concepts at a higher-order category structure, although this more specific finding is not predicted by us. This might be due to the role of emotions acting as a reward cue across different concepts and situations, which might help with concept coherence in a higher-order category structure. However, further evidence would be needed to confirm this interpretation.

Our second hypothesis states that PSVs will have the highest level of concept and category correlations for concrete concepts, followed by intermediate and then abstract concepts. We justify this by suggesting that grounding in object co-occurrences would be more critical for concrete concepts than for abstract concepts, and therefore predict a gradual decline in grounded representations matching those from the LSA space as abstraction increased. Our results directionally support this prediction. LSA correlations, for both concept- and category-level accuracies decrease as concept concreteness decreased. However, it is worth stressing that although not explicitly stated in our hypothesis, we expect a gradual decline in both concept- and category- accuracies across the three concreteness groupings. However, both our LSA match accuracies for categories and concepts show a marked decline for both intermediate and abstract concepts. This suggests that although PSVs are indeed capable of capturing the statistical regularities of objective co-occurrences meaningfully enough for concrete concepts, we overestimated the quality of such semantic representations for intermediate and abstract concepts. However, given that we only have 60 concrete concepts and 40 concepts

across both intermediate and abstract concept groupings, we have to be cautious of generalising our interpretations across the concreteness spectrum. Moreover, even though we have 20 concepts in both the intermediate and abstract categories, given that the concept selection is shaped by LSA neighbours (experimental necessity for meaningful comparisons), this compounds the *small sample bias* with an additional experimental confound, that of *non-random sampling*. Therefore, a useful future investigation would be to extend this comparison across a considerably larger set of concepts, and if sufficiently large, one would expect meaningful categories to occur by chance alone without requiring to “artificially engineer” it in the stimuli as is done in our case with the nearest neighbour criterion.

We predict that intermediate concept correlations with LSA are likely to be in-between those of concrete and abstract concepts’ correlations (hypothesis 3). This hypothesis is only partly supported, because although this is the case for PSVs, it does not apply to scene2vec. Our fourth hypothesis, of scene2vec representations having equal levels of concept and category correlations across the three concreteness groupings, is firmly rejected. Although scene2vec’s respective correlations for intermediate and abstract concepts, within concept and category groupings, are similar, the correlations for concrete concepts remains more than 40% higher for concept-level correlations and circa 10% - 25% higher for category-level accuracies. However, we accept our final hypothesis, scene2vec (compared to PSV) representations have a higher concept- and category-level accuracy. These results reveal that scene2vec representations are superior semantic representations across the concreteness spectrum. Although scene2vec is highly correlated on both concept- and category metrics, it provides a more substantial advantage for representing more abstract concepts than PSVs, given the greater relative differences between the correlations.

Nonetheless, even with scene2vec representations, intermediate and abstract concepts are not as successfully represented as is the case with

concrete concepts. Our results provide both qualitative evidence in the form of coarser differentiation (when present) of more abstract concepts in the hierarchical clustering analysis as well as weaker correlations when compared to concrete concepts. Thus, despite making progress in improving the mechanistic grounding of abstract concepts, with the novel extension of PSVs to incorporate emotions by creating scene2vec, we conclude that our efforts are a modest and incomplete step towards understanding the acquisition and representation of more abstract concepts.

The results of this study also suggest greater interconnectivity between concepts at different ends of the concreteness spectrum, particularly in the case for abstract concepts. In the case of scene2vec, despite more meaningful associations emerging within intermediate and abstract concepts, these are still distinct from the larger set of sixty concrete concepts included in the present study, with some exceptions for concepts like *eye* and *beauty* which share a high degree of semi-automatically extracted object-occurrence cues.

Given our results, this exploratory study provides mixed support for the role of emotions in grounding more abstract concepts successfully. In fact, despite promising results for the semi-automatic extraction of emotions from photos of naturalistic scenes and improvement in the representation of more abstract concepts, this research highlights some of the issues that remain unresolved. Further research questions have been identified to explore the more nuanced relationship between abstract concepts and grounding and other complementary mechanisms. Abstract concepts are a critical aspect of compositionality in human thinking, and we argue that aspects such as generalisability and logical inference are also critical not only for mental operations with abstract concepts but also in constituting the concepts themselves. Similarly, Shallice and Cooper (2013) suggest that there are some fundamental limitations of hub-and-spoke type models with representing abstract concepts, which are also pertinent to our

current computational approach of representing abstract concepts using a simple single-layered neural network.

Shallice and Cooper (2013) use a quintessentially GOFAI/symbolic computational metaphor of an *operator* and an *argument* to re-interpret the sub-symbolic representation of hub-and-spoke network representations, although we think their metaphor can be applied to connectionist models in general, including our models presented in this chapter. In their view, it is reasonable to assume that concrete concepts have a list structure of features, while the operator specifies the type of input representation. Concept representations can be determined by different *isa*, *has* and *can be* relations of spoke properties. In the case of our models, instead of different spokes (e.g. *visual* or *auditory*), one could conceive of different groups of semi-automatically extracted objects and emotions as the features. However, for abstract concepts, Shallice and Cooper (2013) argue that such an approach is ill-suited given the difficulty of deriving a suitable set of features and rules, which is easier for concrete concepts with physical referents in the real world. In fact, according to Shallice and Cooper, a much broader set of nuanced features and operators are likely to be needed for representing abstract concepts. Shea (2018) also suggests that the conceptual content of abstract concepts is likely to be quite distinct from that of concrete words, and reviews the importance of *feelings* and *metacognition*. In our current study, we have only focused on adding emotion expressions to object co-occurrences. Although mechanistically grounding metacognition might be promising, we also claim that doing so could prove difficult and beset with numerous *a priori* assumptions.

The two theoretical positions of Shallice and Cooper (2013) and Shea (2018) might help explain why grounding abstract concepts can be so challenging. However, our results do suggest that emotions are likely to play a crucial constituent part in the content of more abstract concepts. Thus, our conclusion of grounding abstract concepts is not a negative one as is the case with Shallice and Cooper (2013). Further, we also argue that it

might be an intractable task to fully ground all core aspects of semantic cognition across the concreteness spectrum in a single representational hub.

Earlier in this chapter, we have outlined the *embodied abstract semantics* hypothesis of Kousta and colleagues (2011) - how does experiential information such as emotions are important for shaping abstract semantics? Nonetheless, we did not cover some of Kousta et al.'s more extensive arguments around integrating experiential and linguistic information. This omission was motivated by emphasising their core emotion-based hypothesis, which is our focus, and it is also because we aim to mechanistically investigate vital theoretical ideas with the least number of assumptions as opposed to a wide range of flexible assumptions. Therefore, we avoid a highly pluralistic model formulation with poor parsimony. Furthermore, we also avoid incorporating linguistic information given the plethora of theories and models in the extant literature (see Vigliocco & Vinson, 2007).

Nonetheless, Kousta et al. (2011) did caution in their discussion that neuroimaging evidence suggests that there are likely to be more linguistically-relevant brain areas activated in the processing of abstract words. This might further explain why our grounded representations did not perform as well for more abstract concepts. Although given that we are matching the performance against the language-based LSA space, we obviously cannot confound the comparison by including linguistic inputs in this chapter's experiment. Moreover, our objective is to understand the relative strengths and weaknesses of emotional information accounting for semantic representations across the concreteness spectrum.

How can we reliably further our rigorous scientific understanding of semantics from the fields of *artificial intelligence* and *cognitive science*, more broadly and in particular, cognitive modelling? We predict that BrainBench-type litmus tests are increasingly going to play an essential theoretical role in cognitive semantics. Xu et al.'s (2016) BrainBench test suite for investigating the validity of distributed representations stemming

from general machine learning models or, as in our case, more specific cognitive models, will pave the way for more objectively analysing the neurocognitive basis of semantic cognition. Arguably, the current BrainBench stimuli comprising sixty concrete concepts are limiting, but still sufficient for demonstrating the comparability of our *Perceptual Scene Vectors* (PSVs) with language-based distributed models. This is quite promising given that it suggests that our environmentally grounded semantic representations might well be another common format for generating human-like meaning structures. However, only future research containing a significantly larger neuroimaging dataset will be able to support or refute a more generalisable conclusion on the mapping between our computationally derived cognitive semantic representations and those acquired by neuroimaging studies. Future studies will also need to aim for developing more consistent cognitive semantic tests during the acquisition of the neuroimaging experiments in order to provide useful and general benchmarks, given that BrainBench’s neuroimaging benchmarks are obtained from two separate studies with significant methodological differences.

In our present study, our treatment of emotions might be overly simplistic. We include “emotions” or facial expressions in our new cognitive data representation called *scene2vec*. However, we fail to extract meaningful emotion-expressing actions from a small set of naturalistic images automatically. Abstract concepts that are more likely to contain significant proportions of emotional content are also more likely to contain specific types of actions, which might help further differentiate between more general emotion concepts (Shea, 2018). Furthermore, Touroutoglou et al. (2015) found that the basic emotions commonly used in “emotion recognition software” do not correspond with a brain-based set of basic universal emotions. Therefore, facial associations extracted from images do not necessarily constitute emotions *per se*, but *acquired responses* of emotional display. In our stimuli, these learnt responses are likely to be

amplified because online images are probably more likely to be professionally photographed for a particular purpose (e.g. advertising). Thus, our stimuli might contain more meaningful statistical regularities than experienced in everyday surroundings. Another limitation, although somewhat inevitable given our narrow research objective of grounding concepts across the concreteness spectrum using only visual scenes, is the exclusion of linguistic information in our study. Language is a deep resource of the fine-grained statistical regularities for discriminating between broadly similar categories of concepts, which is one of the limitations of our scene2vec cognitive representation. Therefore, now that we have shown exploratory findings in support of grounding concrete and more abstract concepts, to an extent, future studies might want to exceed this “minimum threshold” of inputs considered and opt for more representationally pluralistic approaches.

Borghi and colleagues (2017) postulate a descriptive *multiple representation theory*, which combines embodied and linguistic inputs, such as *situations, introspection, emotions* and *metaphors*. They argue that a single representational framework is not sufficient to account for abstract semantics. Intriguingly, Borghi et al. (2017) do not discuss the need for hierarchical inference mechanisms for the acquisition of abstract concepts, as alluded to by Shallice and Cooper (2013). In our view, opting for simply expanding the input representations is a risky endeavour given the lack of parsimony in resulting theorising due to the underlying complexity and degrees of freedom in the underlying input representations. Therefore, we believe that a *core* set of evidence-based semantic dimensions along with computational mechanisms for hierarchically integrating these components should help develop a more parsimonious theory of semantic cognition and in the process help unify grounded and symbolic perspectives within psychology, which is our objective for chapters 6 and 7.

In closing, our current computational experiments provide support for a weaker version of *embodied abstract semantics*, the hypothesis postulated by Kousta and colleagues (2011). Although our results have shown that emotions can indeed improve semantic representations, especially for intermediate and abstract concepts, the level of explanatory power of visual scene-based grounding also has some limitations given the diminished quality for representing concepts without obvious physical referents. We conjecture that there are likely to be other contributing factors and mechanisms accounting for the content of abstract concepts beyond emotion expressions. Future cognitive modelling would ideally mechanistically investigate a variety of such descriptive accounts in order to understand their relative merits, as initiated in our present study.

Chapter 6

Network Topology of Semantics: Grounding and Relativity of Meaning

6.1 Abstract

Cognitive scientists have a long history of investigating meaning using semantic priming and assuming feature-based representations. However, only recently has there been a focus on the geometrical properties and relations of meaning, led by Binder et al.'s (2016) research on brain-based componential semantic representations and Troche, Crutch and Reilly's (2017) three-dimensional unitary semantic space hypothesis. Both approaches use linear dimensionality reduction to create a low-dimensional Euclidean semantic space from semantic ratings. In the study reported in this chapter, we have collected semantic dimension and importance ratings from 2,062 participants on 544 English words spanning the *concreteness spectrum*. Critically, we also include context-specific

conditions (e.g. *imagine you are cooking*) followed by ratings of the same semantic dimensions. We generate a network topology using non-linear dimensionality reduction (*t-SNE*). Our novel application of graph-theoretical techniques to cognitive semantic networks reveals that (i) semantic networks have a *small-world structure*, (ii) context-free semantic networks are organised lexically on a concreteness gradient, (iii) changes in context can dynamically modulate the lexical network topology, and (iv) *scenes* are the most critical and influential dimension shaping our conceptual networks. Collectively, these findings support a grounded perspective on meaning. *There is no meaning without context.*

6.2 Introduction

How is semantic memory organised? Cognitive scientists often raise this question (e.g. Farah & McClelland, 1991) but a definitive answer remains mostly elusive. One of the foundational aspects of human semantic memory is that detailed recollections of *when*, *where*, *how* or *why* we acquired that information in the first place do not typically accompany its retrieval, a quality sometimes characterised as *noetic* (Postle, 2015). Semantic memory is a hierarchically organised system responsible for the storage, retrieval and processing of facts and concepts (Tulving, 1972). The noetic origins of semantic memory are likely to be the basis for Tulving's (1983) distinction between episodic memories and "other memories" respectively based on whether one phenomenologically experiences the memory retrieval processes as *remembering* (autonoetic consciousness) or merely *knowing* (noetic consciousness). Even though many semantic memories are likely to be initially traced back and grounded in particular episodic instances, more commonly, semantic memories cannot easily isolate the specific influences that initially contributed to the creation and shaping of the memory traces.

Moreover, unlike episodic memories, many of our semantic memories are shared across people in a given culture (Patterson, Nestor, & Rogers, 2007). However, surprisingly little is known about how brain-based activations link to our *cognitive semantic space* - a topological abstraction of all the associations.

Our primary focus in this chapter is the *topography of conceptual space*, which we define as the study of the common geometrical properties, such as *dimensions, shape* and *features* of that space. More specifically, we investigate the topology of complex multi-dimensional meaning, without making reductionist and *a priori* assumptions of linearity. Our emphasis on a complex topology departs from more traditional classifications along single cognitive psychological continuums such as *concreteness*, where words such as APPLE and DOG are on one end of the concreteness spectrum, while other words, such as HAPPINESS and LOVE are on the opposite abstract end. Such a conceptualisation of linear semantic spaces has some implicit assumptions. The two most important ones being (i) the topology is reducible to a single dimension in some meaningful manner, and (ii) the dimension (e.g. concreteness) has psychological relevance in that it facilitates common cognitive semantic processes.

In this chapter, we pursue an alternative approach to semantic topography, preferring to consider multiple cognitively relevant dimensions instead of a unitary one. However, before focusing on the more recent extant literature on creating meaning spaces based on specific dimensions - *conceptual topography*, we will start our review by exploring the origins of studying the core dimensions of semantic memories, dating back to neuropsychological case studies. After all, a conceptual topology, be it linear and straightforward or non-linear and complex, nonetheless is underpinned by one or more basic dimensions. This brief review will lay the foundation for us to explore some of the more recent work on semantic topography and ultimately embark on our original human experimental work on delineating a brain-based model of semantic topology using

psychological measures to examine the structure of the human meaning space.

In the psychological literature on semantics, there are numerous structural perspectives for evaluating the nature of semantic representations (Brugman & Lakoff, 1988; Rosch, 1999; Lakoff & Johnson, 1999; Tyler & Evans, 2003; Lakoff, 2008, 2016; Johnson, 2013). However, we are particularly interested in investigating empirically-based dimensions underlying human conceptual processing. We reason that an appreciation of the vital semantic distinctions will enable us to build an evidence-based semantic topology derived from these primary dimensions. Therefore, to be able to define the space in the first instance and then subsequently study the resultant topology's characteristics and investigate the meaning landscape, we start our investigation with a focus on specific cognitive semantic dimensions.

6.3 Multidimensional Semantic Space

Despite a long tradition of studying semantic cognition and even mechanistically simulating conceptual representations using symbolic and sub-symbolic modelling, investigations of the human semantic space is a very recent endeavour. What do we mean by the phrase *topography of conceptual space*? Inspired by Louwerse's (2011) *symbol interdependency hypothesis*, we believe that our mental concepts are highly interrelated with one another and that this interdependency can be conceptualised in a higher-dimensional space. Since *topography* is the study of the shape and properties of a system (usually physical), theoretical and technical tools from this domain can be used to generate and interrogate the resulting meaning space which emerges from the interrelations of the concepts themselves. The scientific utility of complex semantic topologies, from the perspectives of both cognitive science and artificial intelligence, will be determined by the increased explanatory powers of these spaces compared

to simpler, more traditional semantic maps. Troche, Crutch and Reilly's (2014) pioneering work in uncovering the topology of semantic space starts with a list of 400 abstract and concrete nouns, with ratings across 365 participants and 12 cognitive dimensions - each of which is a survey question that respondents rate. For example, for the dimension *action*, respondents rate each of the target words on a 7-point Likert scale (*Strongly disagree* to *Strongly agree*) for this question: "*I relate this word to actions, doing, performing, and influencing*". After aggregating all ratings across respondents, the dimensionality of the data is reduced to three dimensions using *exploratory factor analysis*. This reduction allows the visualisation of all concepts in a three-dimensional space for qualitative evaluations using a three-dimensional scatterplot and quantification by computing the Euclidean distances between all 400 nouns. The distance measure captures the semantic relatedness of the abstract and concrete nouns. Moreover, Troche et al.'s semantic space's validity is demonstrated by outperforming *Latent Semantic Analysis* (Crutch et al., 2013).

More recently, Troche, Crutch and Reilly (2017) refined their original multidimensional semantic space due to a limitation of Troche et al.'s (2014) *abstract concept feature* (ACF) rating approach. The ACF was the first psychological instrument for evaluating the semantic relatedness of abstract words, which as they state in their more recent work, was biased to capture a dichotomised concreteness space with insufficient and polarised words (see *figure 6.1*). ACF is explicitly biased due to the specific cognitive dimensions selected. However, Troche et al. (2017) remedied this limitation using a set of 14 cognitive dimensions, called the *conceptual feature rating* (CFR) approach. The main difference between the ACF and CFR approach is that the latter includes additional dimensions for capturing sufficient semantic variations for concrete words. Eight out of the original 12 ACF dimensions are carried over to the new CFR, which are as follows: (i) *polarity*, (ii) *emotion*, (iii) *social interaction*, (iv) *morality*, (v) *thought*, (vi) *time*, (vii) *space* and (viii) *quantity*. Four of the original 12 ACF

dimensions are dropped (*sensation*, *action*, *ease of modifying* and *ease of teaching*) in the creation of the CFR, and six new sensorimotor dimensions are added, which are: (i) *visual form*, (ii) *auditory*, (iii) *tactile*, (iv) *olfactory/gustatory*, (v) *visual colour* and (vi) *self-generated motion*. We will briefly outline the core justifications Troche et al. (2017) provide for the inclusion of these dimensions as part of their new CFR instrument.

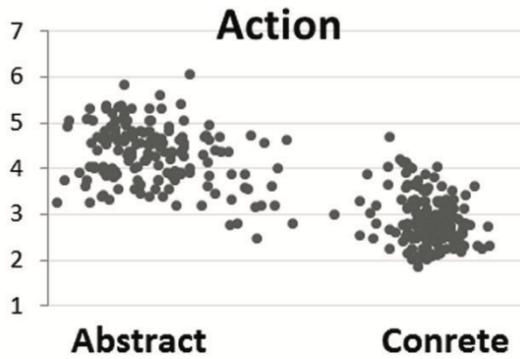


Figure 6.1: Example of Troche et al.'s (2014) dichotomous measure of concreteness, with the x-axis depicting concreteness ratings.

The *polarity* of concepts represents psychological valence (*positive*, *neutral* and *negative*), and is reasoned to be critical for complex goal-directed behaviours. The *emotion* dimension is supported firstly on the basis of research demonstrating the psychological importance of emotional processing in many cognitive activities (Dolan & Vuilleumier, 2003), and secondly, due to recent weak embodiment theories suggesting the role of emotional processing as critical to the conceptual processing of some abstract concepts (Meteyard et al., 2012). Support for the factors *social interaction* and *morality* are respectively rooted in the evolutionary benefit of cooperating with others and neurological evidence from *frontotemporal dementia* (Zahn et al., 2009) showing increased interference with moral concepts. The fifth factor is *thought* and represents higher-order executive functioning for planning and decision making. The dimension *time* is justified because of its role in structuring everyday events, while *space* is justified, somewhat surprisingly, not because of the physical experience of our surroundings but due to the grounded cognition literature's emphasis

on space's central role in metaphoric language (e.g. *love is a journey*). Finally, from the original ACF metric, *quantity* is included due to the prevalence of numerical concepts in language such as *count nouns* (e.g. 12 eggs) and *mass nouns* (e.g. 1 litre of water), which have been shown to be related to both acquisition of concepts and concreteness effects (Gordon, 1985). The new CFR also includes six sensorimotor dimensions to provide greater explanatory power for concrete words, based on Shallice, Warrington and McCarthy's (1983) outline of the importance of sensorimotor aspects for concrete concepts. We show the full dimensions of the CFR in *table 6.1*.

Cognitive dimension	Instruction to respondent
Polarity	"I relate this word to positive or negative feelings in myself."
Thought	"I relate this word to mental activity, ideas, opinions, and judgments."
Emotion	"I relate this word with human emotion."
Interaction	"I relate this word with relationships between people."
Time	"I relate this word with time, order, or duration."
Space	"I relate this word to position, place or direction."
Quantity	"I relate this word to size, amount or scope."
Morality	"I relate this word to morality, rules or any other thing that governs my behaviour"
Visual form	"I relate this word to shapes, forms, textures that I can see with my eyes."
Tactile	" I relate this word to sensations (e.g., texture, shape, temperature) I can feel with my hands or body."
Smell/Taste	"I relate this word to flavours and odours I can smell and/or taste."
Auditory	"I relate this word to sounds, rhythms, etc. that I can hear."
Colour	"I relate this word to colour."
Self-Motion	"I relate this word to my own self-generated movement."

Table 6.1: Troche et al.'s (2017) *conceptual feature ratings* (CFR) instrument, the first column contains the 14 cognitive dimensions and the second column the verbal descriptions for participants.

Troche et al. analyse the 14 cognitive dimensions using exploratory factor analysis to reduce the number of dimensions to three. The three factors are labelled (i) *endogenous* (emotion, polarity, social interaction, morality, motion self-generated and thought), (ii) *exogenous* (colour, smell/taste, tactile, visual form and auditory) and (iii) *magnitude* (space,

quantity and time), and the conceptual space is conveniently plotted in a three-dimensional scatter plot (*see figure 6.2*).

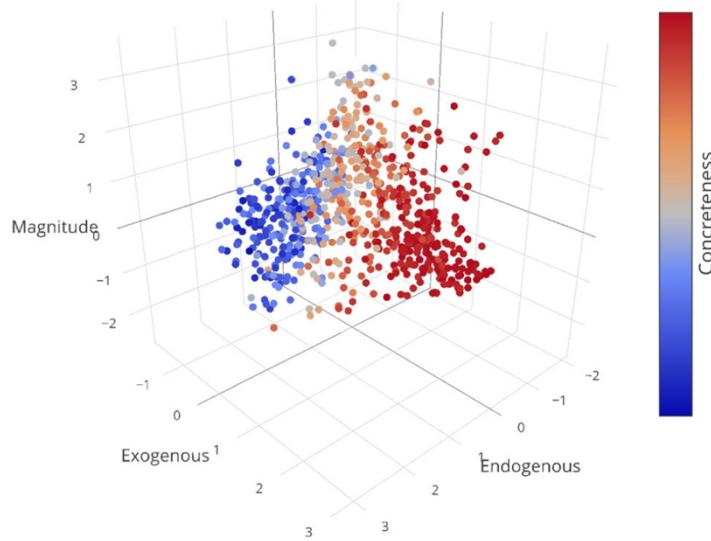


Figure 6.2: The three-dimensional semantic space generated by Troche et al. (2017), where dark red points indicate words with highest concreteness ratings, while dark blue points more abstract words.

Troche and colleagues' refined approach, CFR, reveals an interesting conceptual topology, although focused explicitly on explaining concreteness. Troche et al. demonstrate that there is a continuum between concrete and abstract words, based on the difference between their previous abstract feature ratings (AFR) and the newer CFR approach. They also suggest that due to their conceptual space originating from three psychologically-inspired factors, the resultant multidimensional space is amodal and therefore more naturalistic. We further clarify the second part of this interpretation, which forms part of our discussion. In our view, claiming that the semantic space is amodal supports a moderate to strongly disembodied stance even though one could also argue that if different modalities activate across concepts differentially, then the meaning space itself is multimodal.

Before outlining our specific theoretical motivations for our research, we summarise some of the limitations of Troche et al.'s approach, which may limit the applicability of a CFR-based space in becoming a

general approach to mapping semantics. However, in Troche et al.'s defence, this is not their objective, as they clearly state that their research focuses on the topology of word concreteness. Therefore, many of the limitations identified are more of an opportunity for future research to extend their pioneering work and help continue the development of cognitive semantic analysis in uncovering the conceptual topology.

Firstly, in Troche et al.'s CFR-derived multidimensional space, all semantic distances appear to be roughly equidistant, where concepts are all approximately equally spaced across the MDS spaces. The lack of semantic clusters is surprising. Additionally, we consider that Troche et al.'s focus on creating a semantic map specifically to account for the concreteness spectrum, might have biased the dimensions used. We hypothesise that one of the reasons for their semantic map being poor at highlighting smaller groups of words is due to a lack of sufficiently similar functional and associative words. One can overcome this issue by supplementing the existing set of concepts with additional words from particular conceptual domains, where one would reasonably predict strong associations to exist. However, their choice of linear dimensionality reduction is likely to be a substantial contributing factor to the absence of semantic clusters.

Since Troche et al.'s meaning space itself is created based on latent psychological dimensions, it would be difficult to interpret their results from a computational perspective, given that CFR aims to identify higher-order psychological constructs without mechanistic implications. Nonetheless, if brain-based components motivate the cognitive dimensions, then the emergent conceptual space could be analysed to explore and test predictions based on neuropsychological conditions like *semantic dementia*.

We propose that a conceptual space, constrained by findings from neuroimaging would be a fruitful interface between brain- and cognitive-based semantic spaces. Similarities and differences between these spaces could potentially highlight differences based on hardwired semantic

associations or acquired semantic associations, which might have interesting applications in clinical psychology. Relatedly, this might even shed light on typical and atypical neurological developmental patterns in children and provide early diagnostic signs based on a relatively straightforward CFR-type questionnaire. Disorders such as *obsessive-compulsive disorder* (OCD), are known to be associated with “semantic-morphing” (Bream et al., 2017). A brain-based conceptual space might be able to identify problematic semantic networks early enough for interventions to prevent more severe and debilitating symptoms impacting the daily functioning of individuals - an essential criterion in the *Diagnostic and Statistical Manual of Mental Disorders* (APA, 2013) for many psychological disorders.

Furthermore, the CFR approach implicitly assumes that symbol interdependency of concepts within a space consists of a *non-sparse matrix*, where every concept is related to every other concept, even if the degree of relationship is minimal, due to the Euclidean distances computed across all 750 words. How realistic is this concerning an ecologically valid cognitive data representation? Although we concede that this would be a minor limitation of the current CFR’s conceptual space, as thresholding stronger associations would overcome this, this raises, in our view, a critical question of how to define such a threshold. The main argument we are making is that the statistical and mathematical tools that allow us to define a topology need to be accompanied by cognitively meaningful constraints if the aim is to map the human conceptual space. Therefore, despite that the conceptual space successfully represents concreteness, it cannot account for modality-specific and taxonomically sensitive associations.

There has been a long tradition, since Shallice et al. (1983), of considering different sensorimotor modalities to be more or less critical for particular types of words. How would that be encompassed in the CFR meaning space? Troche et al. (2017) suggest that if their semantic space did indeed correspond to a brain-based semantic space, then they would be

able to transform their statistical model into a powerful computational model and be able to, for example, lesion a portion of the *magnitude* axis in the CFR space (p.12). Even though lesioning a single axis might be viable, we cannot see a precise method for assigning the dimensions a differential set of weights. The CFR-based space currently does not provide relative importance of each of the dimensions. However, overcoming this limitation would allow for more “simulation-like” case studies of semantics from a neuropsychological perspective.

Finally, one last limitation, which arguably in our view is the most important one, is that the CFR-space does not address the effect of contexts on both semantic encoding and retrieval. There is an implicit assumption that there exists a *uniform semantic space*, much like in other distributed models of semantic memory, ranging from older models like LSA (Landauer & Dumais, 1997) and HAL (Lund & Burgess, 1996) to more recent machine-learning based models like *word2vec* (Mikolov et al., 2013). This shortcoming applies to a wide range of semantic topologies and not just the CFR space, but we believe that this limitation prevents the topology from being used in cognitive computational modelling, which would be a good litmus test for any psychologically accurate semantic space. We strongly hypothesise, in our present study, that the semantic space is highly dynamic as a function of context.

6.4 Human Experiment: Mapping our Meaning Space

6.4.1 Visualising Semantic Space

In order to construct an interpretable conceptual space based on fundamental meaning components, we suggest that it is theoretically necessary to use brain-based cognitive dimensions for generating this semantic manifold. The nature of the primitives themselves can change, i.e. could be entirely disembodied amodal feature-type or grounded raw

sensorimotor stimuli, but the problem remains the same. The underlying nature of grounded data is likely to be complex and noisy, but the task of generating a conceptual topology remains mostly the same.

In the case for cognitive scientists investigating semantic spaces, there can be a motivation to not go above three dimensions (reasons outlined in *Appendix B*), for example, in Troche et al.'s (2014) scatter plot of the endogenous, exogenous and magnitude dimensions. However, we argue, that the dimensionality of semantics is very likely to be high-dimensional, which is why Troche and colleagues' three dimensions explain 69% of the variability in the underlying 14 cognitive dimensions, even though our re-analysis shows that five dimensions might actually explain 81%, a further 12% increase in variability explained, which is more consistent with the core dimensions identified by LSA. However, how would we visualise five dimensions simultaneously? Psychologists' answer to this question is typically *multidimensional scaling* (MDS) or *principal component analysis* (PCA).

The realisation that traditional dimensionality reduction algorithms like PCA are suboptimal for representing non-linear manifolds (discussed in *Appendix C*), recently led, in the field of machine learning, to the development of Isomap, by Tenenbaum, De Silva and Langford (2000). Isomap is the first algorithm optimising for global structures by estimating the pair-wise distances in the original space using *geodesic distances*, which is followed by PCA on these distances (Balasubramanian & Schwartz, 2002). Similarly, Roweis and Saul (2000) developed a technique called *locally linear embedding* (LLE), which, for our purposes, is conceptually similar to Isomap, although it centres most of the points at the centre of origin of the map (see Hadid, Kouropeteva, & Pietikainen, 2002). The technique that we advocate is called *t-Distributed Stochastic Neighbour Embedding* (t-SNE), developed by Van der Maaten and Hinton (2008). This technique uses the student's t-distribution to sample nearby points from the original high-dimensional data, which ensures that similar points have

a significantly higher probability of being selected (for distance calculations) compared to dissimilar points. In a second step, t-SNE also computes a similar distribution for the lower-level data representation, which is followed by the third and final step, a reduction in the divergence between the two distributions. Van der Maaten and Hinton demonstrate that t-SNE is significantly superior to both traditional and other more recent algorithms for visualising higher-dimensional datasets, due to t-SNE's ability to optimise for both local and global structures in the data. The nature of the trade-off between local and global optimisation is tuned using the *perplexity* parameter, which we explore in our analysis. This has led to t-SNE replacing most other visualisation methods in many domains of machine learning, such as visualising the hidden layers of neural networks (Mnih et al., 2015), *document search* (Ingram & Munzner, 2015), *recommendation engines* (Shankar et al., 2017) and even *network security* (Kolosnjaji et al., 2016). Recent neuroimaging studies have also used t-SNE for visualising clusters of voxel activations (Mwangi, Soares, & Hasan, 2014).

We hypothesise that given the phenomenological complexity of semantic cognition, it seems plausible to assume that the conceptual topology of the human semantic memory system is highly non-linear. It seems only natural to assume that for explaining a broad spectrum of human meaning, a diverse range of local and global features is necessary for capturing sufficient complexity. Therefore, t-SNE is likely to be well-suited, especially given the perplexity parameter, which, has not yet been investigated in either empirical or computational models of semantic cognition.

6.4.2 Relative Importance of Dimensions

Intuitively, on some level, we can all relate to some of our senses being more central than others for conceptualising diverse things we encounter in our daily lives. A slice of chocolate cake is typically associated

with *visual* and *gustatory* information but not as much with *auditory* sensations. On the other hand, listening to a musician playing on a grand piano is very much associated with *auditory* and *visual* modalities, among others, such as *emotions*, although earlier empirical work has overlooked this cognitive dimension. Lynott and Connell (2013), for example, collected ratings from participants on the extent to which they associated the five primary senses (*seeing, hearing, tasting, touch* and *smelling*) with a range of 400 randomly selected noun concepts in order to create modality exclusivity norms. Their research found evidence for noun concepts being highly multimodal with different modality profiles across noun concepts, even though their focus was limited to the five primary senses (*see figure 6.3*).

In their more recent work (Connell, Lynott, & Carney, 2017), they include *interoception* as a hidden and often forgotten modality critical for the sensorimotor grounding of concepts. Interoception covers a wide range of internal bodily sensations such as temperature, muscle tension, glucose level, hunger and thirst. In their more recent research, they explore the interrelations of interoception with the five primary senses previously studied, with a specific focus on how this manifests itself across the concrete-abstract spectrum. They find that interoception is pervasive across the concreteness continuum, although it is particularly critical for abstract concepts, from which it is the most important for negative emotion words, like *sadness*. Interestingly, the findings also demonstrate that including this additional dimension improves the fit of their model, indicating that additional meaningful information is captured by interoception that cannot be accounted for by the traditional five senses in their original model. Our research addresses one of their original study's limitations of not being able to account for more abstract words.

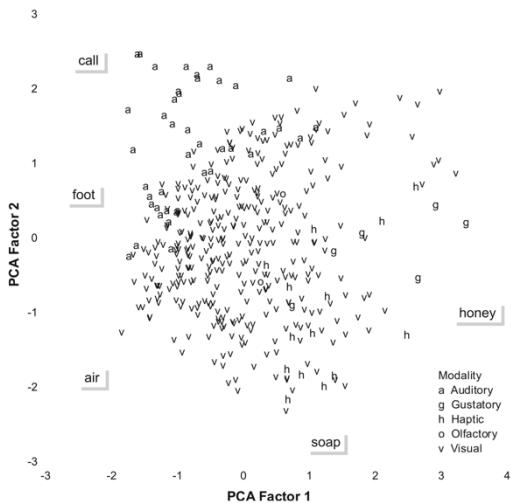


Figure 6.3: Visualising the semantic space generated using PCA of the 400 noun concepts in Lynott and Connell’s (2013) modality exclusivity norm study. The labels correspond to the most dominant modality.

Ultimately, from our perspective of developing a conceptual representation space and investigating this meaning landscape, we build on Lynott and Connell’s (2013) research by focusing on semantic primitives beyond the six senses. Lynott and Connell’s approach is more ecologically valid than using feature sets as the most basic units of semantic cognition, which are quite poor primitives in the sense that features themselves are not ontologically independent lower-level building blocks of meaning. This is also a core characteristic of the research presented in this chapter.

6.4.3 Brain-based Componential Semantics

A radically new conceptualisation of semantic primitives is theoretically proposed and empirically evaluated by Binder and colleagues (2016), with their *brain-based componential semantics* approach. They avoid the pitfalls of “higher-order” semantic primitives in the form of features and instead focus on evidence-based representations based on neuroimaging research highlighting the essential functional divisions of semantics in the human brain. Binder et al. argue that semantically “grounding” the concept BIRD using features such as *has wings*, *has a beak* or *has feathers* is a key limitation of the *standard* feature-based models of semantics. For example, the so-called “feature primitive” *wing* is a higher-

order concept as well, because of links to other nouns like *feathers* or verbs such as *flying*. Therefore, we believe, features being semantically dependent on one another highlights the problem of a symbolic merry-go-round as well as its corresponding combinatorial explosion, which contradicts the well-established evidence of meaning being at least partly grounded in sensorimotor modalities (Meteyard et al., 2012). One of the key common themes running through Binder et al.'s research as well as the present thesis so far is that in order to better understand the process of concept acquisition, we need to pay much closer attention to our everyday surroundings, which Binder et al. term "*experiential attributes*" (p. 2).

Binder and colleagues' research differs from previous research by being a more ecologically valid approach to capturing semantic representations, whereby the emphasis is on common neural correlates associated with independent functionalities. Some connectionist models (e.g. Rogers & McClelland, 2004) purposefully incorporate "features" that are taxonomically relevant to allow the statistical regularities to have both sufficient within- and between-category coherence and incoherence relations, which result in highly interpretable taxonomic discriminations. However, we argue, the modeller-defined features pose a significant *information leakage* risk. This hand-coding of features blurs the boundary of the semantic space as a product of the underlying features or specifically chosen features by the modeller to output specific properties. In contrast, Binder et al.'s view of "macroscopic neural systems that can be distinguished with *in vivo* imaging methods" (p.5) provides a theoretical footing for the inclusion of all the underlying semantic primitives.

Binder and colleagues investigate 535 English words, consisting of 434 nouns, 62 verbs and 39 adjectives. Although the actual queries used to collect the ratings from the participants varies depending on whether a noun, verb or adjective is rated, they are similar to "[t]o what degree do you think of this property as ..." (p.14). These ratings are analysed using a range of descriptive and inferential techniques, spanning simple radial plots and

hierarchical clustering to dimensionality reduction for investigating a total of 65 neural attributes. In order to develop a deeper understanding of the core *independent* cognitive dimensions, Binder et al. performed a data reduction exercise. Sixteen factors emerged from this analysis, among which *Vision/Touch* (*pattern, shape, colour, and texture*) was the most dominant attribute from the factor analysis with an eigenvalue (EV) of 12.81, while the three weakest factors are *time* (EV = 3.22), *luminance* (EV = 2.37), and *slow* (EV = 1.16). We provide a detailed list of all the factors, and how we implement them using the conceptual feature ratings (CFR) approach in *table 6.3*. These 16 factors outperform the more traditional distributed latent semantic representations. Binder et al. visualise the semantic similarities of the representations based on superordinate groupings for making the comparisons more interpretable (see *figure 6.4*). By visually comparing between the brain-based components and the LSA models, clearer subordinate structures emerge, while strong associative relations appear between the various superordinate categories, which is somewhat similar to the results from chapter 3 where our *Perceptual Scene Vectors* (PSV) representations display more fine-grained associations across the concepts. When exploring the results of Binder et al., it is worth noting that they reduced the standard number of LSA vectors from 300 to 65 in order to match their brain-based representation. Interestingly, the results of the 65 and 300 LSA dimensions are comparable.

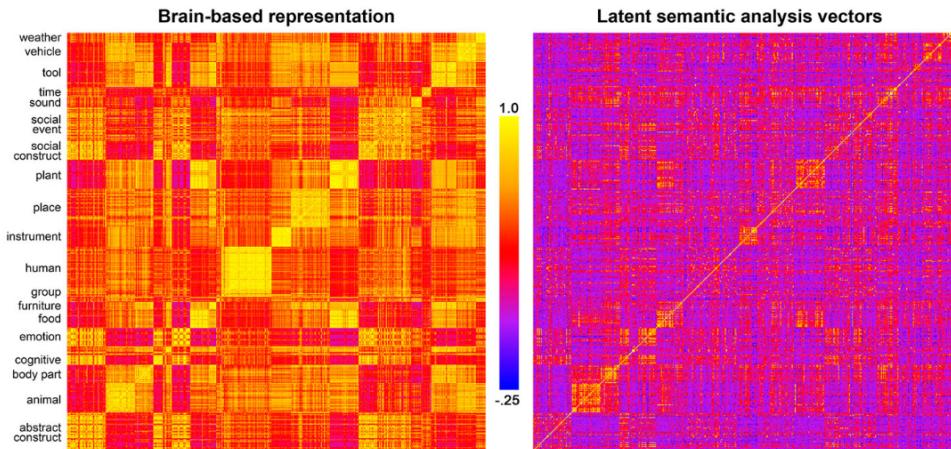


Figure 6.4: Binder et al.’s (2016) cosine-similarity comparison of 434 nouns using brain-based representations (left) with LSA representation (right).

Binder et al. also use Cohen’s d effect size to compare both the brain-based and reduced set of LSA dimensions of within-category versus between-category cosine similarities. At an overall level, across brain-based and LSA dimensions, Cohen’s d is significantly higher for pairs of words in the same category, in comparison to between-category word pairs, although this difference is greater for the brain-based (Cohen’s d mean = 2.64, SD = 1.09) versus LSA representations (Cohen’s d mean = 1.09, SD = 0.85). This difference indicates that even though both the distributional latent semantic and brain-based dimensions can represent the *a priori* category labels successfully, brain-based dimensions are superior at capturing taxonomic-level associations in comparison to LSA dimensions.

Another landmark study is that of Huth et al. (2016), which used fMRI to scan seven participants, while they were listening to more than 2 hours of stories from *The Moth Radio Hour* - broadcasts stories on a wide range of topics. Interestingly, instead of visualising the areas of the brain that are highly active during a specific word appearing in the stories (not ideal given fMRI’s poor temporal resolution), the authors used voxel-wise regression models. From our perspective, it is interesting to note that Huth et al. use natural speech (grounded language) to generate semantic maps for subdividing the human cerebral cortex. Voxel-wise modelling is used because of the complexity of the stimuli that consist of creating a separate

regression model for each of the approximately 60,000 voxels. The researchers note that a core advantage of using voxel-based modelling is the ability to use the model to make predictions on a holdout sample of BOLD responses, which are not included during model training. These models are used to identify the highest predictive responses across 10,470 words, from which the top 20 words form a word cloud for 77 semantic areas in the left hemisphere and 63 for the right hemisphere. This study provides strong evidence for a bi-lateral topology of brain-based semantics, which contradicts previous neuroanatomical evidence suggesting the dominance of the left hemisphere for semantic processing (e.g. Vigneau et al. 2006).

Huth et al. also ran *k-means* to extract a set of data-driven semantic dimensions to map the distinct areas onto a coarse semantic map of the brain, on an inflated map of the cerebral cortex to “blow out” the activations hidden in the sulci of the surface. Despite the limited number of participants, the authors found support for high-levels of inter-participant semantic correspondences. Huth et al. show that semantic maps derived from natural speech are mappable onto the cerebral cortex. Despite this research’s ecologically valid approach to mapping meaning, it implicitly assumes that there is a so-called general semantic map, which plots the meaning of words statically. To clarify, we do not disagree with the presence of significant overlaps in people’s semantic spaces, as that would be difficult to reconcile with the ease and efficiency with which we meaningfully interact in our social and physical world. What we are hypothesising, is that the semantic map itself is not static but always in flux, where experiences gradually alter the semantic network, such that the very act of even accessing our semantic network slightly alters the interconnections based on a number of factors, one of which is the context where meaning encoding or decoding is occurring.

6.4.4 Relativity of Meaning

We postulate, that a central theoretical principle regarding conceptual topologies in particular, but semantics more broadly, is that there is *no meaning without context*. Context-dependent proposals of meaning are not new in either cognitive science or linguistics research, so how does our perspective differ from those previously expressed? We first provide a brief outline of existing theoretical and empirical approaches before detailing our approach.

Schwanenflugel and Shoben's (1983) *Context Availability Model* (CAM) and Paivio's (1991) *Dual Coding Theory* (DCT) are undoubtedly influential precursors to the claim that context influences meaning, especially linguistic meaning. Schwanenflugel, Harnishfeger and Stowe (1988) conducted a study in which they investigated the types of contexts elicited by a range of words, providing support for a positive correlation between the availability of contexts and the ease of recognising particular words. The fact that words can be semantically ambiguous is not a controversial claim in contemporary cognitive science. However, what is important to note is that much of this evidence refers to the ability of linguistic contexts to influence suitable word-level representations, but not necessarily concepts themselves. Yap et al. (2011) showed that contexts facilitate word recognition and words with more semantic neighbours lead to faster lexical decision times in word recognition tasks.

Hoffman, Ralph and Rogers (2013) were the first to create an objective computational metric of linguistic semantic diversity using a corpus-based approach. They build on Adelman, Brown and Quesada's (2006) corpus-based approach, which proposes a metric called *contextual diversity* that computes the sum of all the document in which a particular word occurs in. Hoffman et al. outline that one of the critical limitations with this metric is that it is more likely to be a proxy for word frequency as opposed to genuine contextual diversity, given that the correlation between Adelman et al.'s contextual diversity metric and log word frequency is

higher than +0.95, indicating a nearly perfect correlation between the two measures. This finding is important because it shows that trying to understand the impact of context even in highly structured linguistic data (term-document matrix) can be confounded by factors such as frequency, which obfuscate the validity of measures like contextual diversity concerning semantic representations. The example provided by Hoffman et al. is the word *tax* - it might have an overall high contextual diversity score because it appears in several documents, even though all of these could be related to finance, in which case the actual conceptual diversity would still be low (p.720). In order to overcome this limitation, Hoffman et al. developed a new metric called *semantic diversity* (SemD), also based on corpus data, not by comparing the word frequencies across documents but by generating 1,000-word long context vectors. They reasoned that since the British National Corpus (BNC) consists of 3,125 separate documents of varying lengths, longer documents are less likely to represent discrete topics as they might be entire book chapters or news articles. However, their creation of smaller chunks of documents addressed this inherent limitation of the BNC dataset for context modelling. Hoffman et al. calculated word similarity using context vectors they generated in the LSA space. SemD was a better semantic measure for accounting for semantic judgments in both healthy individuals and patients afflicted with semantic deficits.

Hoffman et al.'s research is a significant first step for semantics research to try to understand the role of contexts on meaning. Moreover, their database of over 30,000 words, tagged with SemD scores, provides useful metadata for related work. However, the research question of how the conceptual space itself is structured was not addressed in this research and has also not been addressed elsewhere with a focus on the relativity of meaning. With *relativity*, we do not just mean words that are *homonyms*, words with different meanings. We also consider words that might share the same sound (*homophones*) or orthographic representation (e.g. "the *bark*

of a tree” versus “the *bark* of dog”) or *polysemous* - where a single word might have two related meaning, as in the case of the word NEWSPAPER, which can be both the organisation or the print/digital publication.

We are proposing that the human conceptual space is gradually constructed based on our experiences across various contexts. Every time we cognitively process a particular concept, the context morphs the concept based on the specific context, task and goals at a given point in time. Of course, some concepts are more fluid than others, and we expect Hoffman et al.’s semantic diversity metric, SemD, to be a good proxy for this. Although Hoffman and colleagues outline a schematic illustration of how context might influence conceptualisation, they do not provide a mechanistic account for doing so based on their semantic features. They provide a diagrammatic/ illustrative example of how the concept PIANO might be more associated with *sound* and *emotion* dimensions in the context of playing the instrument, versus *weight* and *size* if the context is lifting the piano. We propose to empirically test this idea of context-dependent conceptualisation along with the related hypothesis that the meaning space of, for example, a PIANO is not only associated with differentially weighted brain-based dimensions but also is fundamentally different as a function of the context in which it occurs. Contemporary empirical and computational research on semantic cognition commonly overlooks meaning as a context-specific phenomenon.

6.4.5 Objectives

Building on the prior work of Binder et al.’s brain-based semantic components and Troche et al.’s *conceptual feature rating* (CFR) methodology, we aim to synthesise these two approaches to construct a novel semantic network topology. Our research aims to address a number of longstanding and overlooked questions on visualising the landscape of human meaning. Given the high-dimensionality assumption, we argue that only preserving large distances in the conceptual space by minimising the squared error

(like in PCA/MDS) misrepresents the meaning space by overlooking smaller distances. Our first hypothesis is t-SNE will be superior to MDS for representing meaningful and discriminating clusters of concepts. We then use *network analysis* for the visualisation of this data, as this has had promising outcomes in the study of *complex systems*, including *functional brain networks* (Sporns, 2011). We hypothesise that our network visualisations should also aid the interpretation of relationships in the conceptual topology, unlike investigating structure-less MDS plots (e.g. Troche et al., 2017) or word clouds (e.g. Huth et al., 2016). A second hypothesis is that the use of t-SNE will allow us to explore qualitative semantic variations related to local versus global distances. At lower levels of perplexity, concepts within a taxonomy are likely to be more clearly differentiated, due to the focus on smaller distances and vice versa at higher levels of perplexity.

Our third hypothesis is that the conceptual topology will consist of *small-world networks* where concepts are more strongly connected to neighbouring concepts while the vast majority of concepts are disconnected with one another, even though they can be reached with a few links. We predict that the semantic space derived using brain-based cognitive dimensions is going to obey the principle of small-world compositionality, where, on average, the path lengths between the different concepts is relatively short in conjunction with increased clustering within the network (Watts & Strogatz, 1998).

We also predict the presence of taxonomic, associative and functional relationships in our conceptual network topology, which is our fourth hypothesis. However, we caveat that all of these associations are unlikely to be meaningfully represented given the limited number of brain-based dimensions and the large volume of concepts included in our study. Our fifth and probably the most fundamental hypothesis, which is evaluated on three smaller subsets of the concepts, is that the semantic network topology will vary as the background context of a function of

concepts. For methodological reasons, we cannot sensibly shift the context for all words evaluated. However, for three subsets of concepts, we will have a general *neutral-context* situation, along with two *specific contexts* (see *table 6.4*). Therefore, we have a total of six pair-wise comparisons. Our strong prediction is that each of these network topologies will be qualitatively different – that is, semantically organised into meaningful structures for those particular scenarios. In the case of the neutral context condition, we reason that the dominant lexical meaning structure will emerge.

Our sixth hypothesis is that the relative strength of different dimensions will vary across the full range of concepts. Relatedly, our seventh hypothesis postulates that the relative importance of the cognitive dimensions will vary across concepts grounded in different contexts. Our emphasis is on interpretable variations, such that the strength of context-relevant dimensions will increase, while that of context-irrelevant dimensions will decrease. We endeavour not only to understand the relative importance of these dimensions across the concreteness dimension - as Troche and colleagues already researched this - but to investigate whether or not the network topology shifts as a function of differentially activated dimensions. Therefore, our large-scale study will measure the relative importance of the cognitive dimensions for all words. We aim to reveal the natural partitions in our empirically derived conceptual space.

To quantify the relative importance of the dimensions, we adopt a well-established trade-off technique called *Maximum Difference Scaling*, or MaxDiff in short. The MaxDiff technique was initially developed by Louviere (1991, 1992) and then further advanced in a series of subsequent studies (e.g. Louviere, Finn, & Timmermans, 1994; Louviere, Swait, & Andreson, 1995), originally intended for ranking product attributes. MaxDiff is a method for selecting stimuli pairs. The theoretical roots of MaxDiff can be traced back to Daniel McFadden's pioneering work on *discrete choice models* (DCM), summarised in McFadden (1977).

Louviere's (1991) work is an extension of this original DCM theory and methodology for ranking discrete features, in our study, semantic dimensions, that are cognitively difficult to rank order, especially as the number of discrete dimensions to be compared increases. Cognitive scientists have not utilised the robustness and advantages of MaxDiff for ranking cognitive factors. In MaxDiff research, respondents typically only see four or five attributes at a time and select the *best* and *worst* attributes from the selection in front of them. A single MaxDiff experiment might have anywhere between 12 and 16 such individual trade-off screens, where results are aggregated across all participants. From the MaxDiff-derived relative importance metric, we predict that the cumulative importance of all 16 dimensions across all concepts will plateau, indicating that the dimensions are selectively important (hypothesis eight).

In our ninth hypothesis, we predict that higher-order cognitive dimensions such as *human*, *communication*, *self*, *time* and *reward* will have negative correlations with concreteness. Therefore, even though multiple cognitive dimensions are likely to be activated for more abstract words, the dimensions above will be particularly strongly activated. On the other hand, in our tenth hypothesis, we state that lower-order or more perceptually-grounded, cognitive dimensions like *vision*, *ingestion*, *audition*, *place* and *luminance* will be positively correlated with concreteness ratings. We also predict that the semantic network topology will not be solely grouped within the 14 cognitive dimensions (hypothesis 11) because non-linear transformations from tSNE in combination with network visualisation will yield an associatively meaningful network.

Binder and colleagues' research on deriving brain-based semantic primitives is, in our view, a significant step forward for ecologically valid cognitive semantics research, as it provides the scientific community with the opportunity of using these dimensions as building blocks for more general theories of human conceptualisation. We believe that their research provides an opportunity of moving cognitive semantics research beyond

simple and modeller-determined feature sets. Recently there have been concerns in the psychological literature, on the circularity of claims within psychology. Unlike previous reviews of circularity in the psychological literature (e.g. Gigerenzer, 2009), Hahn (2011) provides a balanced perspective on *admissible* and *suboptimal* circularity in both data gathering and theorising. In this review the statement “*God exists because the Bible says so, and the Bible is the word of God*” (p.172) is used to demonstrate how the evidence is reliant on the “self-claim” it is trying to support in the first place. Such classic cases of circularity in reasoning defy falsification despite being logically (deductively) valid, Hahn argues, given that the evidence presupposes the conclusion. However, we are told to not necessarily ignore such assertions as useful given that there are tools for overcoming this, for example, with a Bayesian probability framework which takes into account information about one’s beliefs and adjusts the given probability of a particular hypothesis being true. Inspired by this type of reasoning for empirical research, we argue that the same needs to be the case for cognitive modelling.

There is a proliferation of computational meaning spaces in cognitive science dominated by hand-coded features. For many cognitive modellers, this might be an obvious truism and not worth discussing since simple toy models help us explore the core underlying principles of meaning representation parsimoniously. Let us consider the prototypical case of the concept of ROBIN being part of the superordinate animal category, while PINE being part of the tree category. In most connectionist cognitive models this discrimination is based on the modeller’s assumption that these concepts are different *in a particular way* so therefore should have a vector representing *that* difference.

We claim that this leads to circularity in reasoning when generating computational semantic topologies, akin to Hahn’s (2011) God example, which we reword to apply to our discussion: “*The semantic space exists because the features say so, and the features are the core property of semantics*”.

This type of reasoning, as Hahn outlines, is widespread in science. Although inspired by Hahn’s suggestion of Bayesian inference, we recommend that the fundamental dimensions of semantics should be empirically determined. Avoiding circular definitions ought to enable us to generate a semantic topology independent of our *a priori* assumptions and based on empirically-derived primitives.

Summary of hypotheses

1. Concepts can be discriminated more successfully using the t-SNE dimensionality reduction algorithm in comparison to MDS.
2. Local and global concept associations will be respectively discriminated at lower and higher t-SNE perplexities.
3. The conceptual topology will obey small-world network properties, like sparse density and clustering.
4. The conceptual topology will be organised based on taxonomic, associative and functional relationships.
5. The semantic network of a subset of concepts will morph its structure as a function of context (e.g. context-free, moving house or cooking scenario).
6. The relative strength of cognitive dimensions will vary across network clusters.
7. The relative strength of the cognitive dimensions will vary across concepts grounded in different contexts.
8. The cumulative relative importance of all 16 dimensions will gradually plateau.
9. Higher-order cognitive dimensions such as human, communication, self, time and reward will have a negative correlation with concreteness.
10. Perceptually-grounded cognitive dimensions like vision, ingestion, audition, place and luminance will be positively correlated with concreteness ratings.
11. The semantic network topology will not be structured solely by modality-only distinctions

Table 6.2: A summary of our central hypotheses.

6.4.6 Methodology

6.4.6.1 Participants

The respondents for this study ($N = 2,062$; 44% male / 56% female) were recruited through the UK-based *ResearchNow* panel. They were invited to take part in a Psychology online survey which would take approximately 30 minutes to complete. In return for their participation, they were compensated with panel credits by *ResearchNow*, which can be exchanged for Amazon vouchers. *ResearchNow* is a world-leading professional market and social research panel provider, with over 11 million panellists in their global database spanning 40 countries (<https://www.researchnow.com>). They abide by the regulations of UK industry body MRS (Market Research Society) and the ESOMAR (European Society for Opinion and Marketing Research), and comply with new GDPR (General Data Protection Regulation) requirements. Panellists on *ResearchNow* are carefully vetted using digital fingerprints and an introductory panel survey consisting of 1,000 metrics and geo-IP-validation to ensure they are real human respondents as opposed to bots. The panel has been used in large-scale academic research, such as medical science research (e.g. Graffigna et al., 2015). The data quality has also been evaluated by Schoenherr, Ellram and Tate (2015), who have outlined data quality risks from fatigued respondents, which we take into account during data cleaning. We only recruited participants from the United Kingdom who were native English speakers (self-reported). The only other criterion was that participants had to be able to complete the survey on a computer and not a tablet or mobile device due to the length of the study and the volume of stimuli being presented. The mean age was 45 years (range: 18 - 76).

6.4.6.2 Materials

In total, we selected 500 English nouns drawn from the Medical Research Council (MRC) Psycholinguistic Database (Coltheart, 1981), but included an additional 44 words. All words were pre-evaluated to span the entire concreteness spectrum.

6.4.6.3 Experimental Procedures

In this online study, there were three distinct experimental stages, which respondents completed in the same order. The first stage consisted of the conceptual feature ratings (CFR) using the 16 brain-based dimensions from Binder et al. (2016), where all participants were randomly assigned 80 words (out of the total set of 544 nouns), which were rated on a single dimension randomly assigned to them (out of the total set of 16 cognitive dimensions). Unlike in the Troche et al.'s (2017) experiment, in our study we did not ask each participant to rate all the dimensions or all 544 words in order to drastically reduce the survey length and minimise survey fatigue in respondents given that this is a limitation of online research panels discussed by Schoenherr et al. (2015). Respondents did not switch rating modalities during this stage to avoid interference-based noise artefacts. After the participants read the on-screen instructions, they confirmed their willingness to participate in the study by clicking on the *accept* button.

Participants were instructed that there were no right or wrong answers and that they should use the entire scale and work quickly, yet carefully. Respondents were shown each of the 80 randomly selected words individually (horizontally centred + vertically aligned 1/3 away from the top of the screen) and asked to rate them according to a dimension on a digitised 7-point Likert scale with the following labels: *Strongly Disagree* (1), *Disagree* (2), *Slightly Disagree* (3), *Neutral* (4), *Slightly Agree* (5), *Agree* (6), *Strongly Agree* (7). The numbers in brackets were not shown to respondents

and are indicated here to outline our data codes. The digitised survey Likert scale consisted of a horizontal array of *survey tiles* (see *figure 6.5*). After 136¹⁰ word trials, a “rest window” appeared and asked participants to click on the *next* button to start the second stage of the survey. The scale instructions for each of the 16 brain-based cognitive dimensions are shown in *table 6.3*. The first stage of the experiment, on average, took 8 minutes.

Brain-based cognitive dimension	Instruction to respondent
Vision	I relate this word to shapes, forms, textures that I can see with my eyes
Negative polarity	I relate this word with negative feelings and thoughts, such as sadness, disgust, anger, fear or pain
Communication	I relate this word to communicating with others
Audition	I relate this word to sounds and rhythms that I can hear
Ingestion	I relate this word with tastes and smells
Self	I relate this word to myself and my needs
Motion	I relate this word to all forms of movements
Human	I relate this word to humans, their faces, speeches or other characteristics
Surprise	I relate this word to being surprised
Place	I relate this word to particular places, scenes, settings or landmarks
Upper limb	I relate this word with my arms
Reward	I relate this word with benefits, needs and drives
Positive polarity	I relate this word with positive feelings and thoughts, such as happiness, well-being or arousal
Time	I relate this word to time, duration or numbers
Luminance	I relate this word to visual brightness
Slow	I relate this word to slow things

Table 6.3: Our conceptual feature ratings (CFR) instrument, based on Troche et al.’ (2017) question format of “I relate this word to/with...”. The first column contains the 16 brain-based dimensions derived from the original 65 dimensions evaluated by Binder et al. (2016) and the second column shows the question format for respondents.

¹⁰ This was determined when designing the experiment using the R package *AlgDesign*.

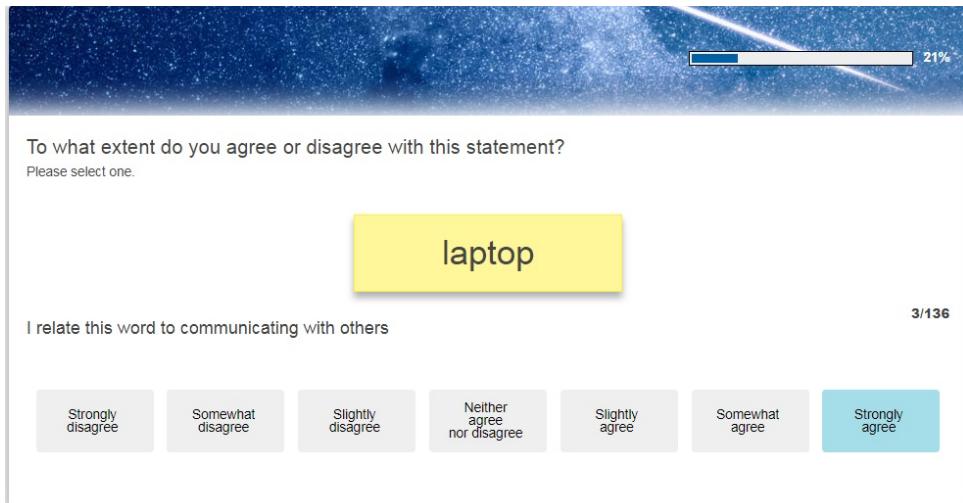


Figure 6.5: A screenshot of how the conceptual feature rating exercise looks like on a computer screen. There is an overall progress bar at the top right-hand side, as well as a fraction count for the specific section of the study.

In the second phase, the MaxDiff study, participants were shown a word and asked to select the most relevant and least relevant dimension from a selection of four dimensions shown on screen (see *figure 6.6*). Each respondent completed 40 MaxDiff exercises. This second stage took an average of 19 minutes to complete. However, for this experimental stage, the order of *words × dimensions* for each respondent, across all their trials was optimally balanced. Our pseudo-orthogonally designed experiment balanced the *number of repeats* (dimensions co-occurring together), *words* and *dimension pairings* and *word order*. Given the large number of words tested, we used the R package *AlgDesign*, which is an algorithmic experimental design package that creates, evaluates and outputs complex experimental designs like MaxDiff experiments (Wheeler, 2014) to optimise the incomplete experimental block design. The final experimental *design file* is used to script the survey in *PsychoPy v3.0*, a high-level Python library for data recording in pre-coded analytical grids to simplify data processing.

Thinking of the word below, please consider the statements and choose one which:

1) MOST relates to the word
and
2) LEAST relates to the word

physics

Click on a button to select.

MOST RELATES		LEAST RELATES
<input type="radio"/>	being surprised	<input type="radio"/>
<input type="radio"/>	humans, their faces, speeches or other characteristics	<input type="radio"/>
<input type="radio"/>	slow things	<input type="radio"/>
<input type="radio"/>	all forms of movements	<input type="radio"/>

Continue »

Figure 6.6: A screenshot of how the MaxDiff exercise looks like on a computer screen.

In the third and final stage of the study, we investigated whether or not the conceptual topology is context-dependent. Respondents were randomly assigned one of six scenarios to read through. Respondents only rated 25 words using the CFR methodology from stage 1. In this between-subjects design (3 pair-wise comparisons), participants completed the ratings once for a single scenario. The six scenarios are shown below, along with the 25 words used. This third stage of the experiment, on average, took 6 minutes to complete.

Context					
1. Moving House	2. Kitchen	3. Fire	4. Water	5. Car Boot Sale	6. Birthday Gift
Please rate the next few words, while imagining that you are about to move house , and need to lift and pack your belongings .	Please rate the next few words, while imagining that you are in your kitchen preparing a meal.	Please rate the next few words, while imagining that you are taking these things from a house on fire .	Please rate the next few words, while imagining things that can float on water .	Please rate the next few words, while imagining things that can be bought in a car boot sale .	Please rate the next few words, while imagining things that can be bought as a birthday gift
refrigerator cool box ice cube tray oven microwave hob kettle toaster spoon fork knife plate chicken pork beef cucumber tomato apple bottle chair cup table door window sink	mobile phone pets children clothes documents heirlooms jewellery money photos laptop TV tablet passport wallet cake chair lamp comb cabbage tangerine plate fork cucumber knife apple			apple banana lemon cranberry cucumber pizza lobster cake food dinner spaghetti diamond jewel jewellery saxophone lamp clothing shirt umbrella corkscrew spatula spider gun mansion puppy	

Table 6.4: Stimuli used in the context-specific experiments (phase 3). Twenty-five words are rated per scenario, with identical concepts split across two scenarios. The neutral-context scenario (not displayed) is derived for these same words from the general phase 1 ratings of all 544 concepts.

6.4.6.4 Data Analysis Procedures

Our initial set of analyses were inspired by Troche et al. (2017), to have sufficient points of comparison between our current study implementing brain-based cognitive dimensions using the CFR method and the original CFR study, which uses *pure* cognitive dimensions.

Following this initial analysis, we build on this with a range of additional analytical techniques, which we outline below.

Before data analysis, we cleaned the data using a set of pre-defined guidelines. However, our criteria were more conservative than those used by Troche et al. (2014). We had a larger sample size and therefore maintained the statistical power of the study despite the list-wise omission of respondents' data. Our online ResearchNow panel was also likely to contain a small cohort of "professional respondents", a problem first outlined by Dennis (2001). We excluded respondents that took less than 15 minutes to complete the entire survey. Perfect *zigzag patterns* on the survey were identified by searching for inter-stimuli (in order of presentation) patterns like "1, 2, 3, 4, 5, 6, 7" or "7, 6, 5, 4, 3, 2, 1" across seven consecutive words.

In order to integrate *concreteness ratings* in our dataset (not collected from participants), we merged our aggregated word-level responses with the concreteness (CNC) score from the extensive *concreteness rating database* (40,000 English word lemmas) provided by Brysbaert et al. (2014). We ran *exploratory factor analysis* (EFA) to evaluate the presence of higher-order factors in our brain-based cognitive dimensions. However, even though Troche et al. (2017) used the commonly suggested *Varimax rotation with Kaiser Normalisation* (Kaiser, 1958), an algorithm that is readily available in SPSS, we opted not to replicate this part of the analysis.

Costello and Osborne (2005) provide an updated set of best-practice guidelines for social science researchers, supporting *maximum likelihood* as the best factor extraction method as long as it is accompanied by analysing the *scree plots* along with a qualitative interpretation of multiple factor test runs. We followed the guidelines outlined in Costello and Osborne (2005). However, like Troche et al. (2017), we also evaluated the inter-rater reliability across our 16 dimensions using *Interclass correlations*.

Our statistical analysis started with non-linear data reduction using *t-distributed stochastic neighbour embedding* (t-SNE). We ran a series of

different t-SNE models with a *perplexity* range of 2 to 181, in order to evaluate the trade-off between the preservation of small neighbourhoods of meaning versus unearthing the global semantic structures. We also used network analysis techniques to graph the conceptual topology.

Finally, we used a simple *Aggregate Logit* model (Lipovetsky & Conklin, 2015) to compute word-level utilities across all dimensions. The MaxDiff data was not analysed using state-of-the-art *discrete choice modelling* approaches such as *Latent Class* or *Hierarchical Bayesian* analysis (Orme, 2009), due to data sparsity. We had a large volume of items rated (544 words across 16 dimensions), despite only having few exposures of each word per respondent.

6.4.7 Results

6.4.7.1 Inter-rater reliability

The inter-rater reliability of each of the 16 cognitive dimensions measured separately on the *conceptual feature ratings* (CFR) and the *Maximum Difference* (MaxDiff) importance scale was analysed using a two-way mixed model. As advocated by Troche et al. (2017), we also used the statistical guidelines from Cicchetti (1994) for interpreting the strength of interclass correlations in psychological testing. According to Cicchetti (1994), interclass correlations below 0.40 are *poor* and between 0.40 and 0.59 the level of significance is *fair*, while the level of significance is *good* between 0.60 and 0.74 and *excellent* when between 0.75 and 1.00 (p. 286). Like Troche et al. (2017), all our dimensions on the CFR rating have excellent inter-rater reliabilities, even though for the MaxDiff ratings of the three dimensions (i) *motion*, (ii) *slow*, and (iii) *upper limb* the reliability was good rather than excellent (see *table 6.5*). This suggests strong consistency in how respondents interpreted and responded to rating the dimensions and the choice-based MaxDiff importance tasks. This is important given that all

following statistics, network analyses and interpretations of our semantic topology are based on this data.

Dimension	Interclass Correlation	
	CFR Rating	MaxDiff Rating
Audition	0.84	0.82
Communication	0.86	0.78
Human	0.85	0.78
Ingestion	0.92	0.87
Luminance	0.83	0.81
Motion	0.83	0.74
Negative polarity	0.86	0.86
Place	0.83	0.80
Positive polarity	0.84	0.78
Reward	0.84	0.78
Self	0.84	0.78
Slow	0.77	0.74
Surprise	0.84	0.78
Time	0.89	0.82
Upper limb	0.84	0.71
Vision	0.83	0.83

Table 6.5: Interclass correlations (ICC) of the 16 cognitive dimensions for both brain-based *conceptual feature ratings* (CFR) and the *Maximum-Difference* (MaxDiff) scores. Interclass correlations that are good, as opposed to excellent, are coloured red.

6.4.7.2 Predicting concreteness ratings with cognitive dimensions

In order to predict concreteness (CNC) ratings, we merged CNC ratings from Brysbaert et al. (2014) with our aggregated word-level ratings from the current study. This resulted in a match-rate¹¹ of 98% (532 / 544). Unmatched concepts ($n = 12$) either were compound concepts (e.g. *ice cube tray*) or had different plural/singular forms. Words with differences in spellings because of British or American English were mapped manually.

14 out of the 16 dimensions significantly predict concreteness ratings. The strongest positive correlation is attributed to *vision* ($r = 0.77$, $p < 0.001$), while the strongest negative correlation to *communication* ($r = -0.54$, $p < 0.001$). Thus, higher ratings on the vision dimension correspond with

¹¹ Match-rate in this context refers to the proportion of words successfully linked between the 544 words used in the present study and Brysbaert et al.'s (2014) database.

larger concreteness ratings, while stronger communication dimension scores are associated with lower concreteness ratings. In line with our predictions (hypotheses 9 and 10), more sensory motor dimensions (*vision*, *ingestion*, *luminance*, *auditory*, and *motion*) are positively associated with concreteness ratings, while higher-order dimensions (*communication*, *human*, *self*, *time* and *reward*) are negatively related to concept concreteness ratings.

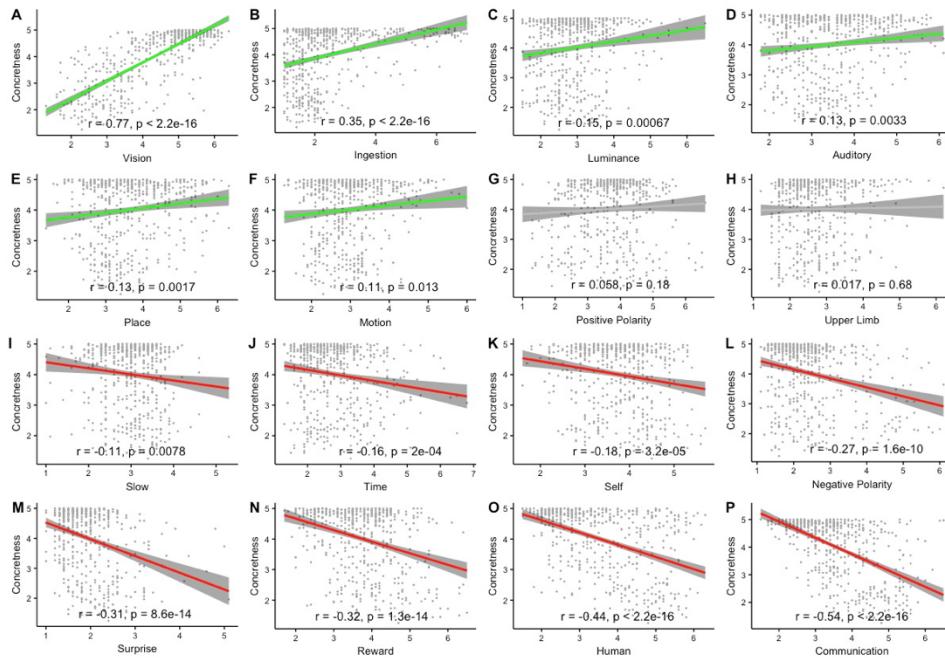


Figure 6.7: Correlations between the 16 cognitive dimensions and the concreteness ratings merged from Brysbaert et al. (2014). Dimensions are rank ordered (from A to P) based on most positive (green) to most negative (red) correlations. The regression line and its confidence interval band along with the correlation coefficients and p-values are shown alongside each graph. The dimensions *positive polarity* and *upper limb* have no significant correlations.

6.4.7.3 Factor Analysis

Troche et al. (2017) use *factor analysis* (FA) to determine the final number of “compound dimensions” for generating the semantic topology. We run a *Maximum Likelihood* FA on the 16 cognitive dimensions (only CFR ratings) based on the guidelines from Costello and Osborne (2005) and Raîche et al.’s (2013) heuristics on non-graphical solutions to selecting the best factors, which yield a 6-factor solution (see *figure 6.8*). However, visually inspecting the *scree plot* reveals a gradual decline in variance

explained as the number of components increases, which supports the independence of our 16 CFR dimensions, given the lack of a distinct “elbow point” in the continuous rate of decline. Based on this exploratory factor analysis, we decide not to aggregate any of our dimensions into higher-order factors. The *factor loadings* (see figure 6.8) even for the first factor, show a wide range of variation (from 0.4 to 0.8). The dimension *reward*, although highest on *factor two*, is also strongly associated with *factors one* and *three* while *factors five* and *six* are single variable factors. Collectively, this does not support a small number of latent variables reliably accounting for the 16 brain-based cognitive dimensions in our study.

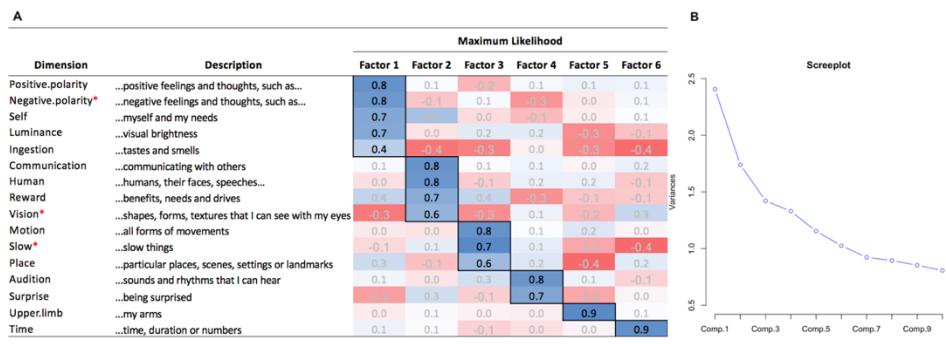


Figure 6.8: (A) Factor Analysis components matrix with conditional formatting (blue for larger numbers and red for smaller numbers). (B) A scree plot of the first nine principal components extracted. Dimensions with a red asterisk, are ones where the scale has been inverted for ease of interpretation.

6.4.7.4 Descriptive Summary

We extract the concepts with the highest ratings across each of the 16 dimensions to assess the *face validity* from an interpretive perspective and to examine the top range of associations on the seven-point Likert scale for evaluating the dimensions. The weakest “top concept” for a particular dimension has a score of 5.1 (concept *miracle* for dimension *surprise*), whereas the strongest is the word *food* for the dimension *ingestion* (see table 6.6), with a score of 7.0. The top concept-dimension pairings are qualitatively meaningful. Larger maximum CFR scores of dimensions are also in-line with greater interrater reliability, from table 6.4, indicating more consistency for concepts on those particular dimensions.

	Audition	Communication		Human		Ingestion	
	Luminance	Motion		Negative polarity		Place	
loud	6.1	mobile phone	6.5	man	6.3	food	7.0
symphony	6.1	telephone	6.4	businessman	6.2	cake	6.9
clang	5.9	debate	6.2	intelligence	6.1	barbecue	6.8
squeak	5.8	announcement	6.0	girl	6.1	broccoli	6.8
piano	5.8	laughter	6.0	sister	6.0	chocolate	6.7
carnival	5.7	negotiate	5.8	boy	6.0	orange	6.6
	Positive polarity	Reward		Self		Slow	
laughter	6.6	honesty	6.5	intelligence	5.7	boredom	5.3
love	6.5	marriage	6.4	knowledge	5.6	grief	4.6
happy	6.5	laughter	6.4	instinct	5.6	walk	4.6
fun	6.3	motive	6.3	trust	5.4	cloud	4.6
affection	6.3	affection	6.3	holiday	5.4	wander	4.4
joy	6.1	development	6.3	family	5.4	flounder	4.2
	Surprise	Time		Upper limb		Vision	
miracle	5.1	time	6.8	hand	6.2	garden	6.2
loud	5.0	duration	6.5	arm	5.8	car	6.2
avalanche	4.8	number	6.5	shoulder	5.3	diamond	6.2
magic	4.6	year	6.3	crossed	5.2	church	6.2
Christmas	4.2	month	6.1	finger	5.2	boy	6.1
fear	4.1	money	6.0	sword	5.0	pool	6.1

Table 6.6: The top-6 concepts (out of 544) for each of the 16 dimensions. The numerical score is the aggregated **conceptual feature rating** (CFR) measured on a 7-point Likert scale.

6.4.7.5 Dimensionality reduction: MDS and t-SNE

Reducing the dimensionality of our semantic ratings is an important step in balancing the underlying signal-to-noise ratio in our dataset and increasing the interpretability. However, from our exploratory factor analysis, we find that a simple linear decomposition into a smaller number of dimensions is not feasible given the lack of an optimal “cut-off” point and a lack of interpretability. Dimensionality reduction aids generalisability and interpretability by extracting key latent interrelations. This is the case for both simple linear- and non-linear dimensionality reduction techniques. However, we hypothesise that the non-linear t-SNE

technique is more suited for semantic data given its machine learning basis in detecting lower-dimensional representations from higher-dimensional manifolds.

We test our first hypothesis on the superiority of t-SNE over MDS, by visualising all 544 concepts using both methods. Our results, in figure 6.9, show that t-SNE is superior at representing concepts globally across the meaning space as well as more locally in associative clusters. On the other hand, the MDS space differentiates extreme items (e.g. *explosion* and *gun*) at the expense of “squashing” all other concepts. In the MDS space, for example, this leads to *sunfish*’s two nearest neighbours being *lamb* and *forest*, while for t-SNE, the neighbours are far more associatively relevant, *pond* and *sea*. These results support our prediction of t-SNE being superior to MDS at meaningfully representing semantic dimensions.

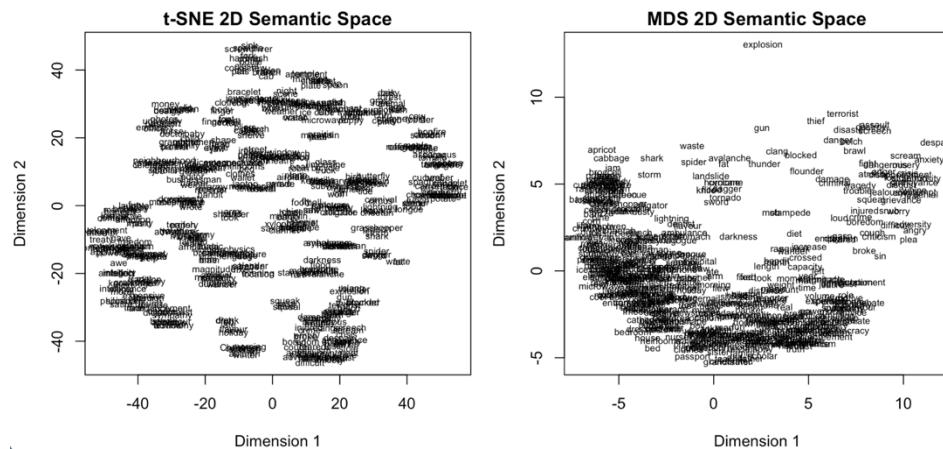


Figure 6.9: Comparing t-SNE’s semantic embedding space (A) with MDS’ space (B) for all 544 concepts. See Appendix D for enlarged images (p. 341).

6.4.7.6 Structure of Network Topology

Our approach to generating a meaning landscape - *semantic topology* - is based on a two-stage process, consisting of (i) non-linear dimensionality reduction and (ii) network visualisation. We use t-SNE, to reduce our 16 CFR dimensions into two dimensions that can capture both non-linear local and global structures within the data. The primary parameter for t-SNE is

perplexity, which can range from two to the *number of cases* divided by three.

In our case, this results in a perplexity ranging from 2 to 181.

The t-SNE algorithm outputs a set of x- and y-axis coordinates, based on non-linear transformations (depending on the perplexity) which we then transpose to calculate the bivariate correlations between all 544 concepts. These concepts form the nodes of the network analysis. Nodes are assigned different colours based on properties such as concreteness ratings. A final association matrix, typically called the *adjacency matrix*, is then compiled based on a threshold applied to each element of the dense correlation matrix (544×544) of pairwise associations. This leads to the generation of a binarised adjacency matrix for plotting an undirected graph, where the vertices lack directions. This is a crucial step, which has significant transformative effects on the underlying network structure. However, the transformation of the association matrix (correlation) to an adjacency matrix (after binarisation) is needed to ensure that meaningful network structures emerge as opposed to having a network where all nodes are interconnected. Different thresholds will produce significantly different network topologies with varying degrees of network sparsity and interconnectivity patterns. Therefore, we explore a range of realistic values and present a small “snapshot” in *figure 6.10*. Realistic thresholds are defined based on assumptions like concepts cannot be mostly interconnected directly, as would be the case with a very low correlation threshold in the region of greater than or equal to 0.50. In our case, both the t-SNE perplexity and the correlation cut-off impact network topology.

Initially, we explore the topology of the network without overlaying the specific concept labels. This allows us to focus on the high-level morphology of the network and understand how this varies as a function of *perplexity* and the *correlation threshold*. We generate ten separate association matrices for every pair of t-SNE perplexity (2, 50, 100, 181) and correlation threshold (0.95, 0.90, 0.80, 0.50). We then average the pair-wise associations for each of the 544 concept-pairs across the ten matrices. The

correlation threshold is then applied to this averaged matrix to produce the binarised adjacency matrix. We visualise this adjacency matrix using network analysis, where the spatial arrangement of the nodes and edges is structured using a spring algorithm. This algorithm minimises the crossing of edges and the spread of nodes to enhance network interpretability (Kamada & Kawai, 1989). This exploratory network analysis reveals high-level structures of the semantic network derived from our brain-based conceptual feature ratings (CFR).

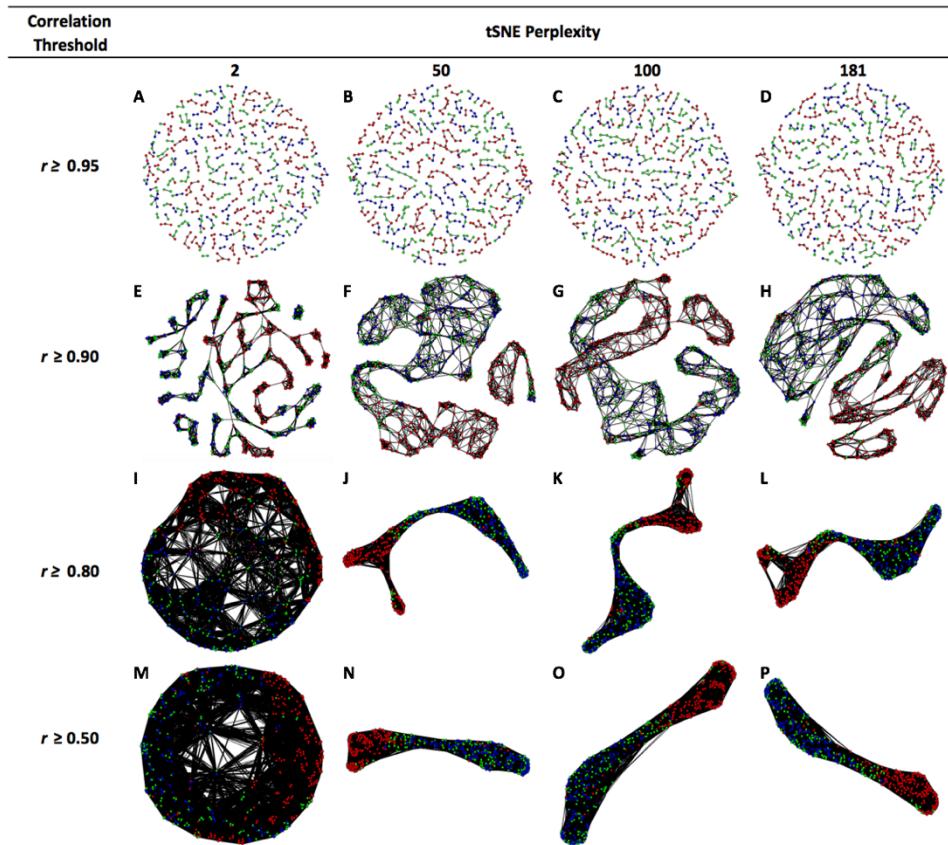


Figure 6.10: A matrix of network topologies as *t-SNE's perplexity* (columns) increases and the *correlation threshold* (rows) decreases, labelled A to P. The concreteness spectrum is indexed using node colour and is binned into the following three bands: *abstract* (red), *intermediate* (green) and *concrete* (blue).

When *t-SNE's perplexity* remains constant (see *figure 6.10*) but the correlation binarisation threshold gradually decreases from greater than or equal to 0.95 to greater than or equal to 0.90, 0.85, and 0.50, a broad global structure in the underlying topology emerges - *concreteness*. More abstract and concrete nodes polarise on either end of the network, with more intermediate-level concepts in the middle. At the lowest level of the

threshold (≥ 0.50), the number of network edges is greatest given that more associations are present in the adjacency matrix. At a global level, the presence of concrete and abstract concepts at two ends of the network indicate that concreteness is a latent structural property of our semantic topology. This finding is particularly surprising in our network structure given that concreteness ratings are not included in the generation of the association or subsequent adjacency matrices.

The network is entirely grounded in the 16 brain-based CFR rating dimensions. Additional metadata such as concreteness ratings are merely overlaid as a network node property for visualisation and interpretation purposes. Therefore, this has no impact on the topology itself. In *figure 6.10*, at the second highest level of correlation thresholding (≥ 0.90), more discernible network structures emerge as predominantly spurious associations are minimised in favour of retaining stronger associations, which are more likely to be semantically relevant. This also enhances the interpretability of the network by balancing the prevalence of false positives and false negatives occurring from the spurious associations. At the same time, the threshold of ≥ 0.90 also captures associations at both a local and global level, unlike at the highest threshold (≥ 0.95), consisting of very short “local chains” of concepts. However, the inclusion of t-SNE’s perplexity as an additional variable adds further complexity, which we wish to highlight before outlining our approach to determining a suitable threshold for both perplexity and correlation extraction for our overall network topology.

At the highest correlation threshold (≥ 0.95), when perplexity increases, the local “cluster chains” increase in length, leading to a slight decrease in the number of components in the network. However, at both the lowest perplexity level of 2 and the correlation threshold of ≥ 0.50 , a large “network ball” emerges due to the high density of network interconnectivity. As t-SNE’s perplexity increases, while the correlation threshold is reduced (moving diagonally from A \rightarrow F \rightarrow K \rightarrow P), a gradual

shift from local to global meaning occurs. However, as the correlation threshold is reduced and the perplexity metric increased, an elongated, dense network structure emerges, with once again, a distinct concreteness spectrum running through the widest diameter of the network. The range of 50 to 181 for perplexity and ≥ 0.90 for the correlation threshold leads to discernible network structures. This is the most useful range we qualitatively identify for generating an interpretable semantic topology.

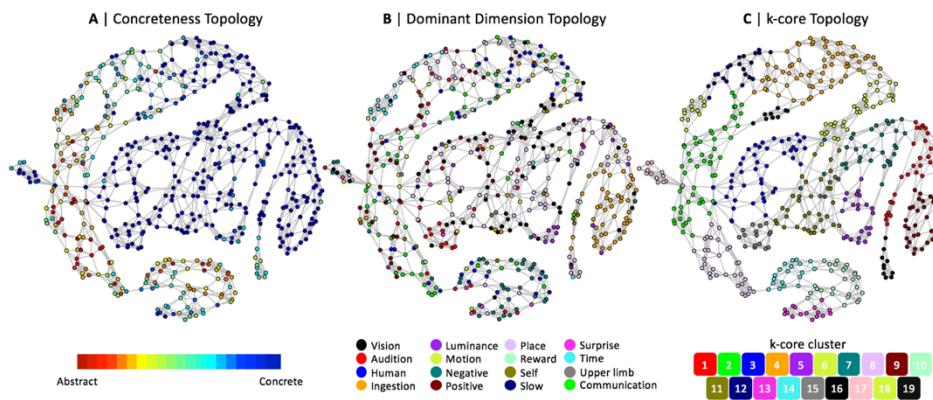


Figure 6.11: Three views of the semantic topology of 544 concepts, based on nodes colour-coded to represent (A) the concreteness spectrum based on ratings integrated from Brysbaert et al. (2014), (B) the most dominant cognitive dimension selected in the MaxDiff discrete choice modelling task, and (C) the k-core clusters.

We evaluate the overall semantic network topology of the 544 concepts based on the (i) concreteness spectrum, (ii) dominant cognitive dimension from the MaxDiff discrete choice exercise and (iii) the k-core clusters. The aggregated concreteness ratings, obtained from Brysbaert et al. (2014), are measured on a Likert scale ranging from 1-5 and overlaid on the topology using conditional formatting (see *figure 6.11a*). The semantic topology reflects clustering of concepts based on concreteness. Moreover, in the upper region of the network, a more continuous spectrum appears from abstract concepts located on the left to intermediate-level concepts located in the middle, and concrete concepts on the right side. The topology, viewed from the concreteness lens, also reveals the absence of highly abstract (red) and highly concrete (blue) concepts adjacent to one another.

The same semantic topology, viewed from the lens of the most dominant semantic dimension (*figure 6.11b*) reveals a slightly different topology than previously reported by Lynott and Connell's (2013) modality exclusivity norm study (*figure 6.3*), in which concepts were neatly grouped based on the dominant modality. We compare the pattern of network nodes with the maximally discriminated cognitive dimension (*figure 6.11b*) and the data-driven *k*-core clustering of the nodes (*figure 6.11c*). The topologies only show some strong network grouping based on cognitive dimensions, predominantly limited to the modality *ingestion*, where the large (orange) network group (*figure 6.11b*) is consistently associated with a single dimension. The cognitive dimension of *negative polarity* is also clustered more locally in the network. Negative polarity is predominantly located in clusters 10, 13 and 14 (*figure 6.11c*). Other cognitive dimensions reveal a highly distributed semantic pattern across the network. In particular, the cognitive dimension *place* (lavender nodes in *figure 6.11b*) seems to be particularly well distributed across the entire network, as it does not fall within any one cluster, but rather, is spread fairly evenly across the network.

Next, we briefly outline the main graph theoretical measures used in network analysis before describing our results (see *Appendix A*, for more details). To the best of our knowledge, since we are the first to generate a semantic network topology using cognitive dimensions, there are no guidelines on which of these measures is most suitable for analysing conceptual networks. Even in relatively more well-established network analysis domains such as *computational neuroscience* and *psychopathology*, there is a lack of agreement (Bullmore & Sporns, 2009).

The most fundamental quantitative network metric is the *degree*, which quantifies the number of connections a specific node has. Nodes with higher degrees can be considered more important to the network. Plotting the frequency distribution of the degrees of all the nodes creates a *degree distribution*, which is useful for analysing network properties. Another

important measure is the *clustering coefficient*, which measures the propensity of a network to have nodes with neighbouring nodes which are more likely to be interconnected to nearby neighbours. Random networks commonly have fewer clusters, whereas complex networks have higher clustering coefficients. A network's *diameter*, is the shortest distance between the two furthest nodes in a network, such that no other two nodes can be further apart as long as the shortest path is taken. We use the *small-world index* (SWI), proposed by Humphries and Gurney (2008) to quantify the extent to which a network adheres to the small-world properties of a network, which in itself is a topic of contemporary disagreement in the network analysis literature (Sporns, 2010). At its core, small-world networks have two key properties; they have high degrees of clustering and short pathways that link these smaller cliques into more global structures. Lastly, another commonly used metric is *network components*, which measures the number of disconnected nodes/node clusters that are present.

In figure 6.12, we depict our exploratory analysis of the relationship between t-SNE's *perplexity* measure and *network components*, *diameter*, *clustering coefficient* and *the small-world index*. However, it is important to stress that these results are only applicable to our CFR dimensions, and are not generalisable to other cognitive semantic networks. Our four correlations show, however, that as perplexity increases, the number of components and the clustering coefficient both decrease and have respective correlation coefficients of $r = -0.88$ ($p < 0.001$) and $r = -0.93$ ($p < 0.001$). This is likely to be the case because as perplexity starts to favour more global associations, the prevalence of smaller specialised clusters decreases leading to fewer components and smaller clustering coefficients.

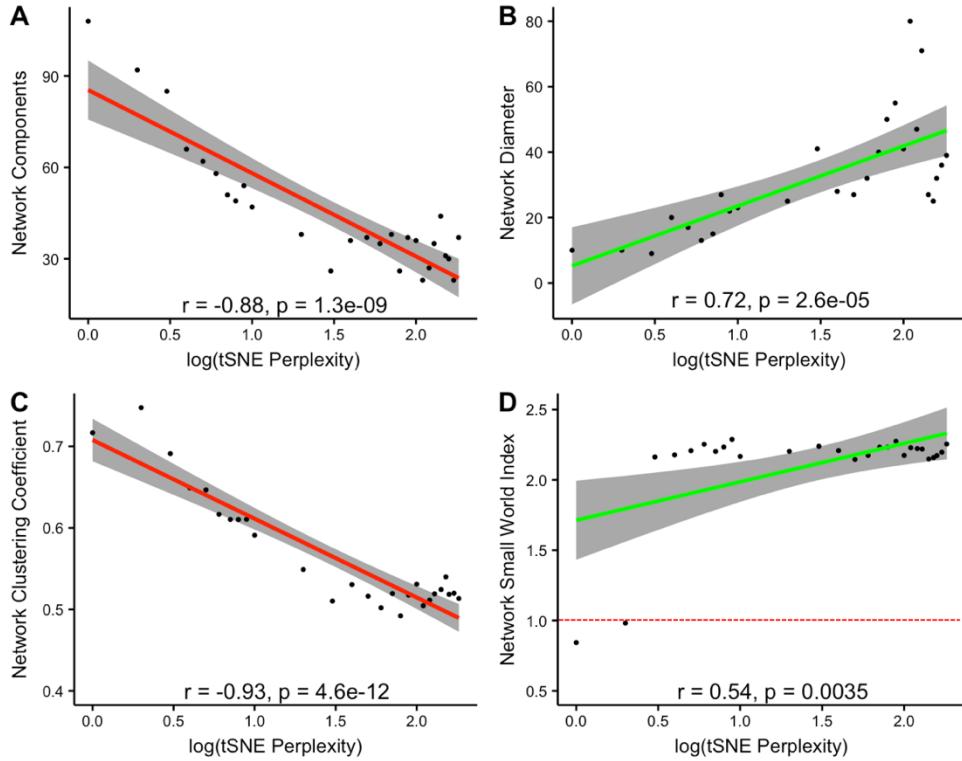


Figure 6.12: Plotting the relationship between a log-rescaled t-SNE perplexity and (A) network components, (B) diameter, (C) clustering coefficient and (D) the small world index (SWI). We include the regression line and confidence interval band along with the correlations coefficients and p-values below each graph.

Predictably, the network diameter positively correlates with t-SNE's perplexity ($r = 0.72, p < 0.001$), which suggests that as more global associations between concepts increase, the network diameter also increases. Lastly, these three relationships also help explain the fourth relation, that of a positive correlation between perplexity and the *small world index* ($r = 0.54, p < 0.001$). The semantic network grounded in cognitive dimensions transitions towards a small-world network as perplexity increases due to fewer disconnected network components and greater clustering of nodes.

6.4.7.7 Visualisation of Network Topology

We now explore the conceptual network topology in greater detail, once again with the help of the concreteness rating lens, but this time, including the actual concept labels. This final network visualisation (see *figure 6.13*) is generated from 50 iterations for each 10×10 combination of t-

SNE *perplexity* (range = 40 to 85, with increments of 5) and *associative threshold* (range = 0.81 to 0.90, with increments of 0.01). A total of 5,000 network models are run during this validation stage, with a *graphical grid search* revealing the perplexity and threshold parameters leading to the most stable range (lowest standard deviation) of *standardised network metrics* (average number of *components*, *diameter*, *clustering coefficient* and *small world index*). Our final parameters are a *threshold* of $r \geq 0.92$ and *perplexity* of 60. Given the size of the network (544 nodes + 3,264 edges), we explore the network structure by starting at the lowest medial portion of the network and gradually moving clockwise. In *figure 6.13*, we qualitatively number the network into *R1* to *R14* regions for referencing purposes.

The most differentiating aspect of the lowest medial portion of the network (*R1*) is the strong community structure, with strong interconnectivity with immediately neighbouring concepts, while being almost cut-off from the wider semantic network. This community is distinctly negative and includes concepts like *tragedy*, *misery*, *jealousy*, *denial*, *fear* and *fight*. This community of concepts also has a diverse range of concepts spanning the concreteness spectrum. However, even the “micro-conceptual” structure within this region, reveals a gradual transitioning from the most negative concrete words at the very bottom of the network (e.g. *gun*, *thief*, *explosion*, *thunder*) to more intermediate words (e.g. *damage*, *criminal*, *tragedy*) and lastly more abstract words (*contempt* and *disgust*).

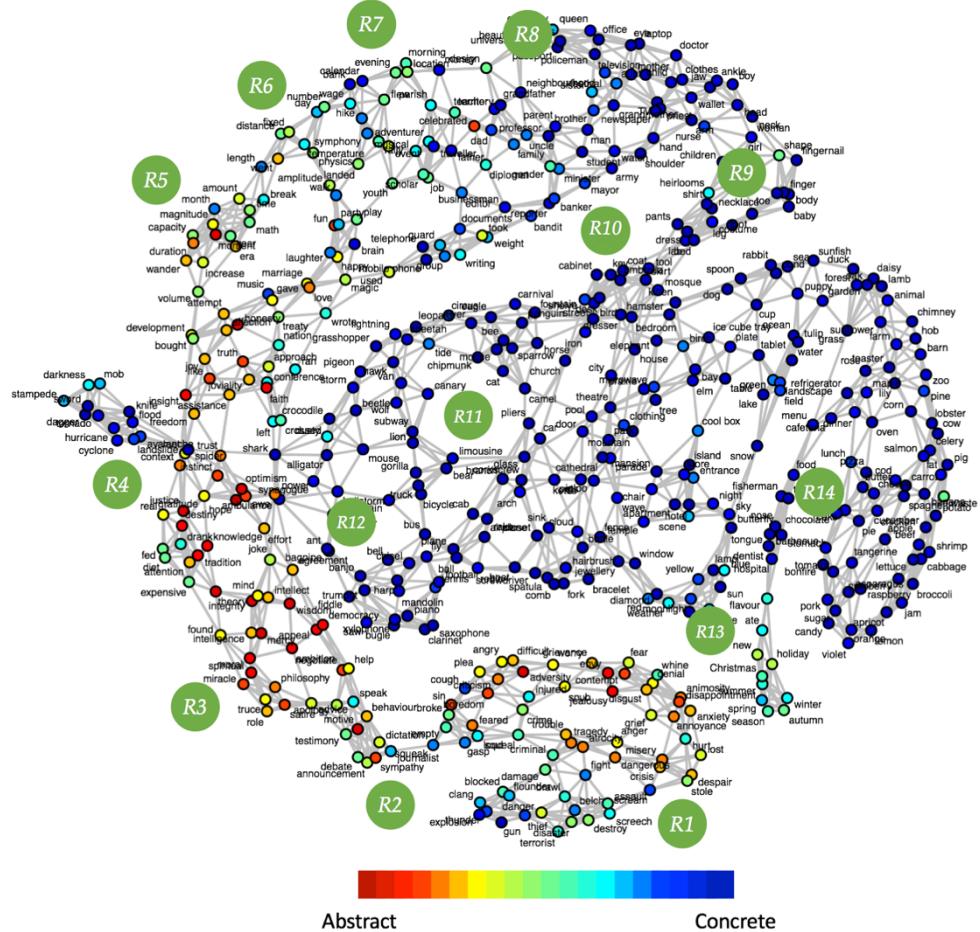


Figure 6.13: Overall semantic topology, including concept labels and the nodes of the network colour-coded according to the concreteness spectrum, ranging from **red** (abstract) to **green** (intermediate) and **blue** (concrete). The green numbered circles (i.e. R1...R14) highlight different portions of the network to aid discussion of more specific network neighbourhoods. See *Appendix D* for enlarged images (p. 342 - 343).

In region *R2*, there is an interesting *network bridging phenomenon*, wherein the negative community of concepts are connected to the main portion of abstract concepts via words like *journalist*, *dictation*, *sympathy* and *debate*. These concepts are intermediately concrete words, which then connect to the most abstract portion of the network. This next region (*R3*) consists of the most abstract concepts within the semantic network and is defined by words such as *wisdom*, *philosophy*, *intellect*, *spiritual*, *intelligence* and *mind*. *R4* is a highly distinct region of the network, given that most concepts are concrete and yet it is positioned away from the “main blue body” of concrete concepts and adjacent to the more abstract words. This small cluster predominantly consists of *natural disaster/danger* concepts, like *avalanche*, *landslide*, *cyclone*, *hurricane*, *tornado* and *flood*, which in turn are

adjacent to the words *dagger*, *knife* and *sword*, which finally lead to *stampede*, *darkness* and *mob*. We find these microstructures within a small cluster of concepts interesting because it reveals that t-SNE’s non-linear dimensionality reduction, can indeed lead to global structures like this cluster of disaster emerging, while still maintaining local micro-structure within the cluster. Another example of this is the link between this set of concepts in *R4* and the most substantial portion of concrete words from regions *R9-R14* being connected by the path from *avalanche* → *shark* → *alligator*, which then fans out to other less dangerous animals and subsequently vehicles and other concrete concepts in *R11*.

Further moving clockwise through the network topology, leading to *R5*, we find a small number of concepts related to the theme *magnitude*, with words such as *magnitude*, *capacity*, *math*, *duration*, *capacity*, *amount*, *moment*, *era* and *volume*. This region consists of a range of intermediate to abstract words, transitioning to *R6*, with more related intermediate-level concepts like *length*, *distance* and *day* leading to *R7* with concrete words like *calendar*, *evening*, *morning* and *location*. This region also has other concepts like *bank* and *wage* as well as *event*, *musical*, *adventure* and *rally* near one another. Collectively, traversing between regions, *R5* → *R6* → *R7* also features a *concreteness gradient* from greater abstraction to more concrete concepts. The next region (*R8*) is a fairly broad region with less of a theme other than the first central region where the network is entirely concrete. Despite that this region contains a mixture of various concepts, they are related. There are strong micro-groupings of concepts that can be meaningfully interpreted. Towards the “top” of the region, *professional* and *place* concepts are consistently grouped, respectively containing some of these words: *professor*, *doctor*, *queen* and *policeman* as well as *university*, *office* and *army*. Concepts like *minister*, *mayor*, *banker* and *bandit* are immediate neighbours. Region *R8* includes *body parts* like *hand*, *shoulder*, *ankle*, *jaw*, *hand* and *neck*, all arranged reasonably tightly but also *family relations* such as the concepts *parent*, *grandfather* and *dad*. Relatedly, in the next region (*R9*), the

concepts *women*, *girl* and *children* appear, although the region is more representatively summarised by the dual themes of *body parts* (e.g. *body*, *leg*, *finger*, *fingernail*, *neck* and *head*) and *clothing* (*pants*, *shirt*, *dress* and *costume*). Once again, despite that more global structures bind all of the concepts in the similar regions of the semantic topology, more local structures are simultaneously maintained by the direct connectivity of strongly associated relations like *fingernail/nail*, *neck/head* or *leg/pants*. This pattern is repeated in *R10*, with *cabinet* and *key* being adjacent concepts, and this region being more loosely defined by *homeware*, with other concepts like *dresser*, *iron*, and *bedroom* being nearby. However, this particular portion of the network also seems to have significantly more spurious associations with animals and non-related words like *mosque*. Region *R11* contains two main classes of concepts, namely *animals* (e.g. *cat*, *mouse*, *camel*, *canary*, *beetle*, *wolf*, *chipmunk*, *lion*, *bear* and *gorilla*) and *vehicles* (e.g. *limousine*, *truck*, *train*, *cab*, *bicycle* and *plane*).

An example of a strongly clustered and meaningful group of concepts can be seen in region *R12*, which comprises *musical instruments* (e.g. *piano*, *mandolin*, *trumpet*, *harp*, *fiddle* and *xylophone*), which are connected to “loud” concepts like *bell* and *plane* before other concrete concepts, less discriminated by sounds appear. Similarly, *R13* is also a fairly discrete cluster of *seasonal* concepts like *Christmas*, *holiday*, *winter*, *spring*, *summer*, *autumn* and *season*. Lastly, *R14* is characterised by a wide range of *food/animal* concepts (e.g. *lunch*, *meal*, *pizza*, *pork*, *cabbage* and *shrimps*), transitioning upwards to *animal keeping* concepts (e.g. *zoo*, *farm*, and *barn*) and finally *flowers* (e.g. *daisy*, *tulip* and *sunflower*). Although *R14* contains numerous animal concepts like *pig* and *cow*, the regions do not include animals not associated with food (e.g. *bear*), and concepts like *rabbit* are more on the periphery of the food regions, closer to prototypical pets like *dogs*, *kittens* and *hamsters*.

6.4.7.8 Small-worldness of Topology

This semantic network topology is evaluated formally on its small world properties using Humphries and Gurney's (2008) *small-world index* (SWI) metric. Although detailed network benchmarks are still limited, and non-existent for psychologically-derived semantic networks given the absence of such models in the extant literature, the recommendation is that a small-world index of greater than one is the *lenient threshold* for a small-world network, whereas the more *conservative boundary* is three. We calculated the small world index for our semantic network, with a *threshold* of $r \geq 0.92$ and *perplexity* of 60 to be: $SWI_{SemNet} = 10.17$. Since, to the best of our knowledge, there are no widely accepted approaches for evaluating the statistical significance of network parameters, we follow the approach outlined by Sporns (2010), and created a *null distribution model* consisting of a *randomly* derived adjacency matrix, with identical number of nodes (544) and edges (3,264).

Our approach, however, differs slightly from merely sampling a Gaussian distribution with a numerical range fitting our *conceptual feature ratings* (1-7 Likert scale) and then generating the association matrix using this random data. The reason is that in order to maintain the same number of edges after selecting our $r \geq 0.92$ threshold, we also need to have the same distribution and set of correlation coefficients in the matrix. Otherwise, the number of edges in the network topology would change, which becomes a confounding variable in our comparison. Therefore, we generate our random association matrix by random sampling (within a cognitive dimension) from our real association matrix (without replacement). This way we ensure that the associations between the concepts and the dimensions are random, but the actual values in the adjacency matrix have the identical distribution to our real adjacency matrix. Our *null distribution arbitrary topology* is shown in figure 6.14.

Comparing our semantic network (*SemNet*) to the *null distribution model* (*RandNet*), on graph-theoretical metrics, provides support for

distinguishing our semantic network from a random network structure. SemNet has a diameter of 56, while RandNet's diameter is only 7, which means SemNet's shortest path between its two furthest nodes is eight times that of RandNet's. SemNet also has a clustering coefficient magnitudes apart from RandNet ($SemNet_{CC} = 0.5628$, $RandNet_{CC} = 0.0095$), indicating that our semantic topology has significantly more tightly knit groups. Lastly and most importantly, the small world index of this random network is $SWI_{RandNet} = 0.9991$, well below the more conservative benchmark of three and the more lenient threshold of one, for qualifying as a small-world network. Comparing our SWI_{SemNet} with $SWI_{RandNet}$ supports our cognitive semantic topology consisting of a small-world network architecture. Our semantic network topology demonstrates an increased propensity to form clusters with paths connecting different clusters of concepts, leading to a small-world network architecture. These results are supported by *bootstrapping* the network metrics through 100 iterations of the null model and calculating 95% *confidence intervals* (see *table 6.7*).

	SemNet	RandNet (N = 100)			
		Mean	Mean	LCI (95%)	UCI (95%)
Components	10	2.44	2.179	2.701	
Diameter	56	6.76	6.67	6.85	
Clustering Coefficient	0.5628	0.00953	0.0094	0.00966	
Small World Index	10.17	0.9991	0.9886	1.0096	

Table 6.7: Network metrics for SemNet and bootstrapped RandNet, including 95% confidence intervals for the null model (LCI: lower confidence interval, UCI: Upper confidence interval).

Null Model: Random Topology

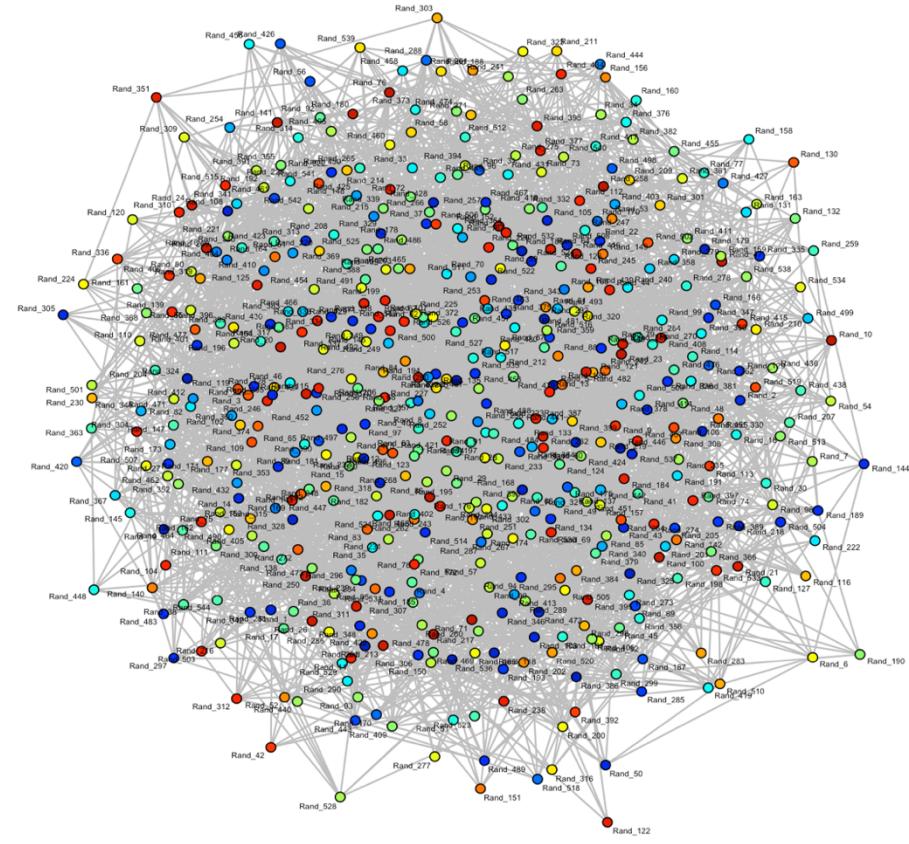


Figure 6.14: A random topology based on shuffling the semantic dimensions across 16 dimensions. This random network topology is generated using the identical set of parameters as our *real* semantic topology. This *null distribution model* also has 544 nodes, 3,264 edges, an *association threshold* ≥ 0.92 , and *t-SNE perplexity* = 60. The node colours are based on the original concreteness spectrum data and as such, is random in this topology. See Appendix D for enlarged images (p. 344 - 345).

6.4.7.9 Importance of Dimensions

Our final analysis objective at the *overall network level* consists of formally analysing whether or not all cognitive dimensions are equally distributed across the entire network. Based on our earlier qualitative evaluations, we found some tentative support for the *place* dimension being more distributed across the network, while the *ingestion* dimension is more localised in the food network cluster. Since we are not aware of a formal statistical or mathematical analysis method for analysing the “distributional equality” from a graph-theoretical perspective, we have repurposed a well-known economic indicator of inequality - the *Gini coefficient*, to meet our analytical objective. At its core, the Gini coefficient

measures the inequality among values in a density or frequency distribution. Gini (1912) proposed this metric as a measure of accurately determining the *wealth distribution* of nations. A coefficient of 0 (no concentration) indicates perfect equality as all values are equally distributed (e.g. everyone in the country has the same wealth), whereas a coefficient of close to 1, indicates perfect inequality, for example, one person has all the wealth. In our analysis, we use the standard computational formulation of the Gini coefficient (Cowell, 2000), in which we replace the distinct populations with our 19 network clusters, and the *wealth measure* with the number of nodes in a given cognitive dimension that are maximally activated from the MaxDiff ratings. The Gini coefficient is operationalised by generating an *incidence matrix*, with 16 rows (one per dimension) and 19 columns (one per network k-cores cluster), capturing the distribution of concept nodes maximally discriminated across the network clusters (see *table 6.8*).

	k-core Cluster																		
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19
Audition	0	3	6	2	0	0	1	2	0	0	3	2	3	2	11	0	1	1	0
Communication	1	12	0	13	0	3	1	13	0	2	2	0	0	2	0	0	0	0	3
Human	0	3	0	18	0	5	0	5	0	5	0	1	0	5	0	0	0	0	2
Ingestion	11	2	1	1	0	2	1	0	36	0	3	0	0	0	0	2	0	0	0
Luminance	3	0	2	0	6	0	2	1	2	0	0	0	0	0	0	1	0	0	0
Motion	0	3	6	3	0	2	0	0	0	0	1	2	1	1	1	0	2	1	0
Negative.polarity	0	1	0	1	1	0	0	1	1	17	0	0	9	6	0	0	3	0	2
Place	11	2	13	8	10	2	14	0	0	1	10	8	1	1	0	5	2	1	0
Positive.polarity	0	19	4	6	0	1	2	7	0	0	3	1	0	0	0	1	0	0	0
Reward	1	6	0	4	0	3	0	1	0	0	1	1	0	0	0	0	0	0	1
Self	3	2	0	3	1	2	0	2	0	0	1	0	0	1	0	0	0	0	0
Slow	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
Surprise	0	1	1	0	1	0	0	0	0	1	0	0	2	1	1	0	1	0	0
Time	0	0	0	1	0	0	1	0	0	0	0	6	0	0	0	1	0	9	0
Upper.limb	0	1	0	2	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1
Vision	2	0	10	2	4	13	12	0	3	1	12	1	0	0	3	0	4	0	0

Table 6.8: The incidence matrix for calculating Gini coefficients.

In addition to calculating the Gini coefficient, we also evaluate two additional measures. Firstly, we compute an adjusted *Lorenz curve* (see Dorfman, 1979, for details). This is plotted based on rank ordering the incidence matrix (for each cognitive dimension) using the number of network clusters activated by a given dimension, ranging from 0 to 1. On the y-axis, we have the *proportion of network clusters* cumulatively activated,

while the x-axis is defined by the actual number of *ordered network clusters*, ranging from 0-19 (see *figure 6.15*). To illustrate this with a specific example, consider the *place* dimension. From the incidence matrix in *table 6.8*, we can ascertain that 15 out of 19 clusters have a non-zero incidence, meaning 4 out of 19 clusters are not activated. Therefore, in *figure 6.15A*, *place* (dark blue curve) starts to increase from point 5 on the x-axis. Furthermore, we use *hierarchical agglomerative clustering* (with *Ward's distance*) to group more similar cognitive dimensions based on their Gini coefficient, to avoid plotting 16 curves, some with minimal differences. The green dashed line in *figure 6.15A* is the “line of equality” and represents the ideal cumulative network contribution if a dimension were to activate all clusters equally. Thus, the closer a cognitive dimension’s curve is to this dashed green line, the *more distributed* it is regarding overall network importance.

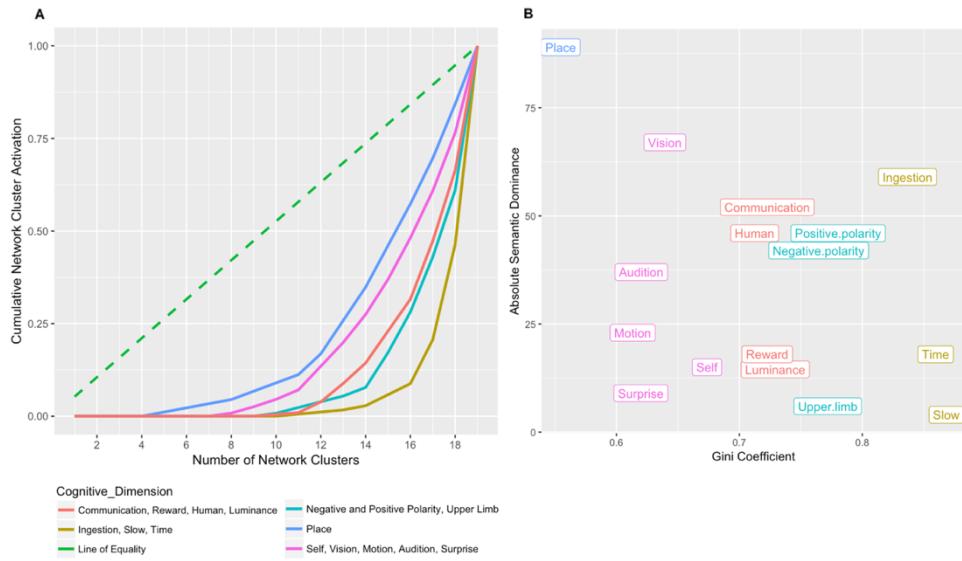


Figure 6.15: (A) An adapted *Lorenz-style* plot of the cumulative distribution of network dimensions’ maximal activations across the 19 k-core network clusters. (B) An overview of the 16 cognitive dimensions plotted based on the Gini coefficient and the *absolute semantic dominance* - a measure of the number of nodes being maximally activated. In both plots A and B, similar cognitive dimensions are grouped for ease of visualisation.

In *figure 6.15B*, the second measure we include is the *absolute semantic dominance*, which is the sum of all the node incidences across 19 cluster for a given cognitive dimension (row sum of incidence matrix). The reason we include this is that the Gini coefficient, as well as our adapted Lorenze-style curves, are relative measures as opposed to absolute ones. In

other words, if a single hypothetical dimension only activated a single node in every cluster, its Gini coefficient would be 0.00 (perfectly equal), even though 19 out of a total of 544 nodes were activating less than 4% of all the concepts in the network. Similarly, we calculate, that if a dimension only activated 1 node, then that dimension's Gini coefficient would be 0.95 (extremely unequal). Therefore, we suggest that the Gini coefficient is contextualised using our *absolute semantic dominance* measure to indicate the overall importance to the network.

Collectively, all of these metrics converge to one of our study's main findings, namely that the cognitive dimension *place* is the most important contributor to our semantic topology. The dimension *place* activates the greatest number of network clusters (Gini coefficient, $G_c = 0.55$), and also maximally activates the greatest number of overall network concepts (89 nodes). This is followed by four groups of cognitive dimensions with similar Gini coefficients, in ascending order (more to less important to the network):

group_1 = {self, vision, motion, audition, surprise}: *average*(G_c) = 0.63,
group_2 = {communication, reward, human, luminance}: *average*(G_c) = 0.72,
group_3 = {negative polarity, positive polarity, upper limb}: *average*(G_c) = 0.77,
group_4 = {ingestion, slow, time}: *average*(G_c) = 0.85.

Additionally, based on *figure 6.15B*, we can also see that both *vision* and *ingestion* also activate a large number of nodes. In the case for *vision*, this is not as interesting, as it happens to be more distributed on both the Gini coefficient and higher in absolute concept node activations. However, this is not the case for the dimension *ingestion*, which is positioned in group 4, with the lowest equality score (highest Gini coefficient) and the steepest cumulative distribution curve, but yet based on the absolute number of concept nodes activated, it is ranked third, behind *vision* and *place*.

In *figure 6.16*, we visualise this discrepancy between *place* and *ingestion*, as both have a large number of concept nodes activated, 89 and

59, respectively, but the distribution across the entire network is skewed for *ingestion*, with strong preferential activations only for clusters nine and one, hence the significantly higher Gini coefficient. From these metrics and visualisations, we start to develop a deeper understanding of the rich and complex interconnectivity between the particular basic componential semantic dimensions and resulting network topology.

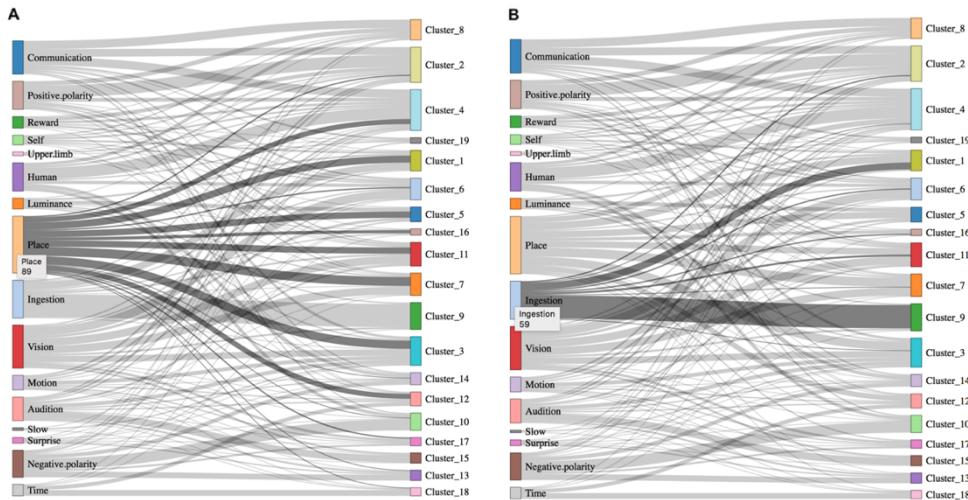


Figure 6.16: (A) Sankey plot of the incidence matrix with *place*-to-cluster relations highlighted. (B) Sankey plot of the incidence matrix with *ingestion*-to-cluster relations highlighted. Even though both dimensions have a large number of concept nodes activated, the spread for dimension *place* is greater.

So far, we have established a range of findings based on our semantic network topology, ranging from the small-world structure, the role of perplexity in highlighting different aspects of local versus global relations in human semantic networks, and that some cognitive dimensions, such as *place*, are more distributed and important. Next, we turn our attention to evaluating particular sub-sets of our network in order to investigate the relationship between context and semantic topology.

6.4.7.10 Context-dependent Network Topology

In this section, we review three separate (between-subjects) network comparisons, all with the common objective of understanding whether context shapes semantic network topology. The three comparisons are similar in format, with each consisting of 25 concepts from the main

experimental phase 1 CFR ratings, which is our *neutral context*. Two specific contexts follow this general condition. All networks within the three comparison sets have an equal *number of nodes* (25 concepts), same *correlation thresholds* (≥ 0.75) and t-SNE *perplexity* (8) but the number of edges can naturally vary based on the distribution of coefficients in the adjacency matrix. In these comparisons, we do not keep the number of edges constant because we are running pair-wise comparisons between networks as opposed to a *null distribution* model.

In our first comparison between kitchen-related concepts (*figure 6.17*), a visual inspection of the *neutral-context* condition and *home move* as well as *cooking* context reveals strong structural differences. A pair-wise comparison of the *home move* vs *general* and the *cooking* vs *general* networks reveals that in both contexts, there is increased clustering of concepts as well as a difference in the number of network components. All network metrics are summarised in *table 6.9*. The degree distributions (*figure 6.17D*) are particularly different for the cooking network, which is bi-modal, compared to the other, slightly positively skewed unimodal distributions. In the home move network, presumably, concepts not typically associated with moving (e.g. *sink*, *window*, and *door*) appear to be clustering together, while edible items also form an independent component. In the cooking context, the edible items form a distinct cluster but are connected to a highly dense cluster, which is also reflected in the relative difference in the *network density* with the cooking context scoring 3x higher than in the two other conditions. Lastly, the *network diameter* of the general context is circa twice the size of the other two networks (see *table 6.9* for the metrics).

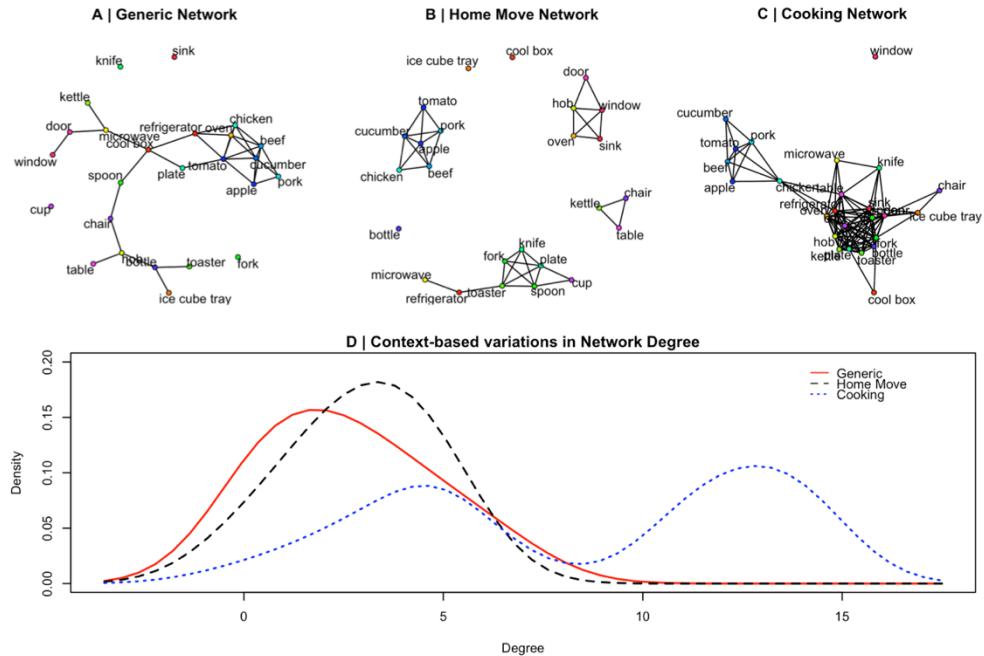


Figure 6.17: (A) Semantic network for neutral context, (B) home move context, (C) cooking context and (D) showing a comparison of network degree distributions using density functions.

In our second comparison between the general condition, a *house on fire* and a *water buoyancy* scenario with a different set of concepts (figure 6.18), like in our first comparison, the neutral condition's concepts are structured associatively. However, the two context networks are markedly different in both their visual structure and their corresponding network metrics (table 6.9). The degree distributions for both the neutral and fire contexts are non-Gaussian. In the neutral network (figure 6.18A) the concepts are divided into five components (three of which are single concepts), where the large components are split between *food* and *household* and *people-related* concepts. Local structures are also meaningfully interpretable, for example, *wallet* and *money* are adjoining as are *laptop* and *mobile phone*. In the *house fire* condition, there are two highly dense groupings, which we presume to be strongly related to a lower cluster of items “not worth rescuing” and an upper cluster of “things to rescue”. Interestingly, the concept *TV* bridges these clusters, which we interpret as a concept that falls in-between the two clusters semantically, as it could belong to either one of these clusters. Similarly, distant concepts in the

neutral condition (e.g. *mobile phone* and *pets*) that have a diameter of five only have a diameter of one in the fire network indicating the stark difference in conceptualisations.

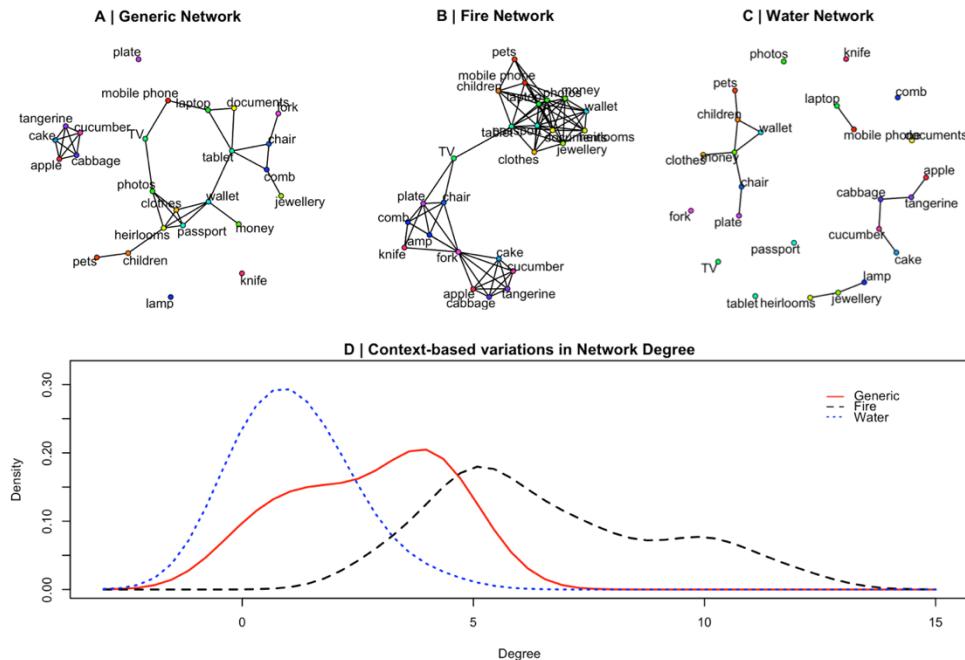


Figure 6.18: (A) Semantic network for neutral context, (B) house on fire context, (C) water buoyancy context and (D) showing a comparison of network degree distributions using density functions.

The *water buoyancy network* (figure 6.18C) is harder to interpret. The network topology, low network density, increased number of components, and a lower clustering coefficient suggest a different meaning space than in the two other conditions. We speculate that the two large components respectively consisting of *food* and *pets, children, some valuables* and a *chair* might be “more buoyant” components, although *plate* is also part of one of these components, which weakens our speculative interpretation.

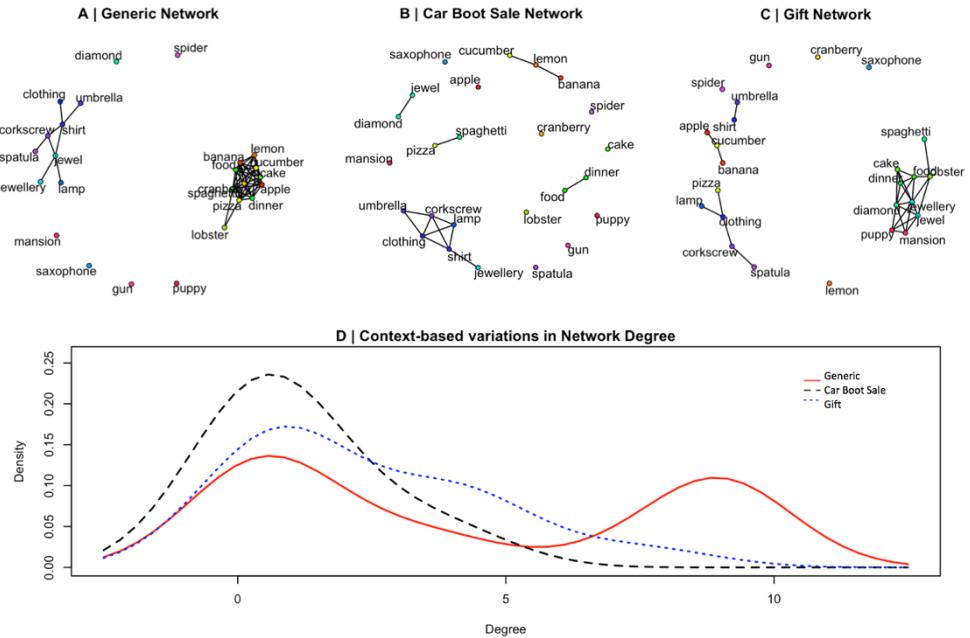


Figure 6.19: (A) Semantic network in neutral context, (B) *car boot sale* context, (C) *gifting* context and (D) showing a comparison of network degree distributions using density functions.

Lastly, in the comparison of our third set of concepts (figure 6.19) between the neutral context and two “commercial” contexts (*car boot sale* vs *gifting*) the first commonality we find interesting is that in the neutral-context scenario, both the *network density* and the *clustering coefficients* are very high. Once again, like in the previous *fire* context, in both the commercial contexts, strongly associatively relevant clustering of concepts occurs. In the *car boot sale* scenario, items like *umbrella*, *corkscrew*, *shirt*, *clothing* and *lamp* are tightly grouped (≥ 2 edges), while *jewellery* is connected to this component, via *shirt*, by a single edge. This component seems to be consistent with the types of objects one might typically come across at a British car boot sale. Similarly, the gift network has a densely interconnected cluster of items we would consider suitable objects or animals as gifts. This “gift” cluster is also meaningfully clustered locally with food on one side (*food*, *lobster*, *spaghetti*, *dinner* and *cake*) and objects and animals on the other (*diamond*, *jewellery*, *jewel*, *puppy* and *mansion*). Atypical concepts form isolated components across all three networks, which include the words *gun*, *saxophone* and *spider*.

		Size	Density	Components	Diameter	Clustering Coefficient	Small World Index
Context Set 1	Generic	25	0.11	5	9	0.543	2.77
	Home Move	25	0.12	7	4	0.779	2.27
	Cooking	25	0.35	2	5	0.788	1.30
Context Set 2	Generic	25	0.11	5	6	0.626	2.84
	Fire	25	0.28	1	6	0.711	1.50
	Water	25	0.05	12	4	0.200	3.13
Context Set 3	Generic	25	0.18	8	2	0.887	1.01
	Car Boot Sale	25	0.05	15	3	0.652	3.98
	Gift	25	0.10	9	3	0.635	1.50

Table 6.9: Key network metrics across three context comparison sets, each containing a neutral context and two context-specific networks.

Collectively, these three comparison sets provide strong support for a *dynamic* and *context-specific* grounded perspective on semantic topology. We have so far qualitatively reasoned why these network structures might vary. We now illustrate a more quantitative approach to interpreting the topological network differences. From the aggregated brain-based *conceptual feature ratings* (CFR) data, we calculate an original metric for understanding the underlying reason for *why* one semantic topology differs from another. In order to do this, we create the *differential activation quotient* (DAQ), which calculates two separate quotients, one for each context, by computing the *average importance* of a single dimension over the relevant subset of words compared to all other dimensions over the same subset of words. Once two quotients are calculated, one is subtracted from another, arriving at the DAQ (see *equation 6.1*). Each cognitive dimension has a DAQ, which leads to a total of 16 differential quotients per comparison. We do not take the absolute difference because we are interested in the direction of the difference. For example, in *equation 6.1*, we calculate the DAQ for the first comparison *neutral* (context 1) versus *home move* (context 2) for both the *upper limb* and *vision* dimensions. We can see that the DAQ for the *upper limb* is +35%, which means that this dimension shows an increase of 35% in the *home move* context compared to in the neutral condition. Whereas, for *vision*, the DAQ is -32%, indicating that, on average, the visual dimension is less important for the *home move* context.

$$DAQ = \Delta Dim(C_1, C_2) = \left(\frac{C2_{Dim} = \frac{1}{n} \sum_{i=1}^n x_i}{C2_{All Dim} = \frac{1}{N} \sum_{i=1}^N x_i} - \frac{C1_{Dim} = \frac{1}{n} \sum_{i=1}^n x_i}{C1_{All Dim} = \frac{1}{N} \sum_{i=1}^N x_i} \right) 100$$

$$\Delta UpperLimb(C_{generic}, C_{home move}) = \left(\frac{C2_{UL} = 4.33}{C2_{All Dim} = 3.77} - \frac{C1_{UL} = 2.44}{C1_{All Dim} = 3.04} \right) 100 = +35\%$$

$$\Delta Vision(C_{generic}, C_{home move}) = \left(\frac{C2_V = 5.24}{C2_{All Dim} = 3.77} - \frac{C1_V = 5.19}{C1_{All Dim} = 3.04} \right) 100 = -32\%$$

Equation 6.1: The *differential activation quotient* (DAQ), is the difference in a single dimension across two contexts and is based on the CFR ratings of the cognitive dimensions across these contexts (C_1, C_2). In our study, we are only interested in the pair-wise comparisons between the different contexts and the neutral context.

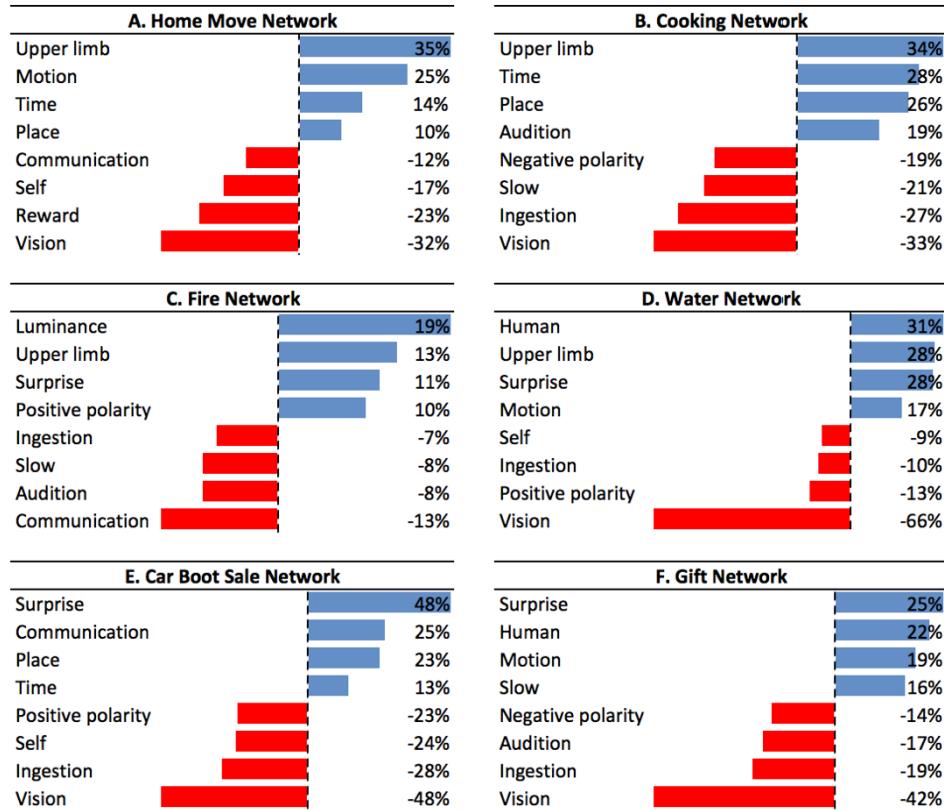


Table 6.10: Summary of the *differential activation quotient* (DAQ) between the six grounded scenarios and the neutral context. We only show the top and bottom four dimensions.

Given that the semantic network topology is grounded in these ratings, the DAQ provides a simple route through which we can understand why the semantic topologies are different based on variations between a grounded scenario and the neutral-context situation. In *table 6.10*

we provide a summary of the highest and lowest DAQs for the six distinct context comparisons. All the DAQs in *table 6.10* are compared with their respective neutral condition. Next, we selectively explore the DAQ findings from two out of six comparisons.

In the *home move* condition, dimensions *upper limb* and *motion* becoming more important, while *reward* and *vision* are less vital for conceptualising these particular concepts in the home move “frame of mind”. In the *cooking* condition, once again *upper limb*, but this time, *time* and *place* are also more important, and somewhat unexpectedly *vision* and *ingestion* are less important. However, that might be because the instructions directed the participants to imagine cooking as opposed to tasting the food. This is more likely to be grounded in our sense of time/space and less on the more dominant visual and ingestion dimensions. This also would be in line with *ingestion* and *vision* being more important in the “default” lexical semantic route of conceptualising the 25 kitchen- and food-related concepts. The topology of meaning is dynamic as a result of the underlying primitive components changing in importance across contexts.

Now turning to the *fire* network, the DAQ results suggest that the dimension *luminance* is strongly activated in the fire condition, probably not because of the task of having to take things out of the burning house, but simply because fire is associated with brightness, which is captured in the CFR dimension *luminance*. Unsurprisingly, the *upper limb* dimension is also more strongly activated, which aligns with the context of having to engage in physical activity. Dimensions *surprise* and *positive polarity* also increase, most likely because a house on fire is a startling event and when thinking about saving valuable belongings, pets and children, this might elicit stronger emotional ratings. Regarding dimensions that are less activated, *ingestion* and *slow* are considerably weaker, which is likely because of the presence of food concepts in this subset of 25 words, which might have had to be suppressed in the *fire* context, given its irrelevance in an immediate

emergency context. Similarly, *slow* is another dimension that would be seen as less important in the context of having to save one's belongings from a house on fire, which elicits the idea of quick actions.

6.5 Discussion

Our investigation of the human semantic topology using a large-scale normative study of brain-based cognitive semantic dimensions via t-SNE and network analysis reveals new insights, such as cognitive semantic networks obey the principles of *small-world interconnectivity*. Most importantly, we provide evidence contradicting long-standing "static theories" of semantics in cognitive science (e.g. PDP), linguistics (e.g. LSA), and the fields of artificial intelligence and machine learning (e.g. word2vec). However, in addition to our exploratory findings, we also support numerous previously established findings, which inspired this current work, as well as raise questions regarding existing interpretations of static semantic spaces. In this section, we first examine the results supporting or contradicting previous results and explanations, before turning our focus to the original contributions of this study. Next, we outline some core limitations of the present study, and where feasible, suggest possible improvements for overcoming these. Finally, we conclude with a discussion of potential implications specifically for *cognitive modelling* of semantics and more generally for *cognitive science* and *artificial intelligence*.

6.5.1 Contributions

Our adaptation of Troche et al.'s (2017) conceptual feature ratings (CFR) using Binder et al.'s (2016) 16 brain-based components reveal very high levels of inter-rater reliability, comparable to those obtained by Troche et al. for their original 14 cognitive dimensions. Our results suggest that not only is it methodologically feasible to use neuroimaging-based semantic

components in cognitive psychological research but also that the brain-based semantic dimensions are phenomenologically meaningful. Similarly, for the MaxDiff ratings, we demonstrate the feasibility of gathering data on the importance of semantic dimensions using discrete choice methods from econometrics. Cognitive rating approaches and econometric techniques are rarely integrated, but our results demonstrate compatibility. High inter-rater reliability supports greater cross-over between neuroscience and cognitive science methods when investigating semantic representations. Motivated by Troche et al.'s (2017) discussion, in the context of clinical applications, our research demonstrates the viability of using cognitive semantic ratings derived from neuroscientific dimensions, which might unlock the possibility of developing empirical and computational models of meaning for "synthetic lesioning" of specific dimensions to reflect neurological damage and observe cognitive semantic deficits. High levels of inter-rater reliabilities across our brain-based cognitive dimensions support such research opportunities.

This study also demonstrates a range of different relationships between 14 out of the 16 concreteness (CNC) ratings. In line with previously reported findings from both Troche et al. (2014) and Troche et al. (2017), concreteness ratings positively correlate with more sensorimotor semantic dimensions (hypothesis 10). As ratings on specific dimensions increase (*vision, ingestion, luminance, auditory, place* and *motion*) so do concreteness ratings. This relationship can be explained by concrete concepts being more likely to have real-world referents with reliable variations across sensorimotor dimensions. Similarly, higher-order semantic dimensions like *communication, human, reward, surprise* and *negative polarity* have strong negative correlations with concreteness ratings (hypothesis 9). Our results lend further support to empirically evaluating semantics using a componential lens as pioneered initially by Crutch et al. (2013), given the replicability of well-established findings on concreteness ratings.

Our results support a concreteness gradient in semantic networks. The concreteness dimension is a latent factor in the structuring of the semantic topology for concepts, which, we find to be limited to the neutral context scenarios. This finding supports the notion that decontextualised concepts are more likely to be associated with lexical similarity ratings capturing surface-level associations as opposed to deeper associative relations (Barsalou & Wiemer-Hastings, 2005). Furthermore, this is also in line with the finding that more abstract cognitive dimensions like *human* display strong negative correlations with concreteness in neutral contexts.

However, in contrast to the findings reported by Troche and colleagues, our semantic dimensions are not reducible to the previously established *exogenous*, *endogenous* and *magnitude* factors. Although Troche et al. claim that their 3-dimensional conceptual space originating from psychologically-inspired factors is amodal, our results suggest otherwise. Based on evaluating the cumulative variance explained by each cognitive dimension, we do not find graphical evidence for a distinctive “elbow” in the number of core dimensions. Non-graphical methods for extraction of an optimal number of components also do not yield psychologically interpretable dimensions. Lastly, evaluating the concepts with the highest dimension-loadings across all 16 cognitive dimensions reveal strong, meaningful variations in the types of concepts across all dimensions. Together, these descriptive and exploratory analyses support a multidimensional perspective, where a three-dimensional model fails to represent the complexity of our conceptual associations. Non-linear methods for dimensionality reduction and representation needs to gain traction in the study of cognitive semantics. Intuitively, this makes sense given the highly complex nature of semantics, which, arguably, are the core “building blocks” of all human meaning-making.

Our comparison of a simple two-dimensional plot using both classical MDS and non-linear t-SNE reveals that the semantic space is more meaningfully clustered using the latter technique (hypothesis 1). As

predicted, the semantic space generated using MDS discriminates large differences between concepts exceptionally well (e.g. *terrorist*) while “squashing” the semantic space for concepts with smaller but important differences. Further support comes from our qualitative explorations of quality of both MDS and t-SNE spaces. The interpretability of t-SNE’s embedding space is also superior to that derived from MDS which returns irrelevant nearest neighbours like SHARK-CABBAGE. We, informally, interpret this as a result of MDS “optimising” for global distances as opposed to local ones, which leads to concepts with large differential ratings plotted in disproportionately distant parts of the embedding space. This “extreme plotting” is problematic for the generation of semantic topologies, because of the sub-optimal interpretability of the concepts. Therefore, we use the t-SNE transformed dimensions as the input to our network visualisation and analysis given the preservation of local and global semantic relations (hypothesis 2).

Despite the metaphor of “semantic networks” spanning decades in cognitive science research (e.g. Reitman, Grove, & Shoup, 1964; Brachman, 1977; Bower, 1970), to the best of our knowledge, we are the first to use cognitive ratings to surface these interrelationships using network analysis. Secondly, the network topology allows us to investigate new associative measures (e.g. *clustering coefficient*), which helps uncover hidden underlying patterns of interconnectivity within cognitive semantic networks. Future research, perhaps from a developmental perspective, could investigate the growth of clustering in semantic representations throughout early child development, which might provide benchmarks for typical and atypical developmental pathways. We provide initial evidence in support of cognitive semantic topologies obeying the properties of small-world networks (hypothesis 3), as previously shown for *biological neural networks* (Sporns & Zwi, 2004) and *language* (Cancho & Solé, 2001), but not for cognitive semantics. Small-world networks, with their unique properties of a combination of short and long path lengths, increased

clustering and greater hierarchical abstraction might be crucial for investigating the topology of semantic networks. Our exploratory results of varying t-SNE's *perplexity* levels (from more local to global) demonstrate a recursive property within semantic topologies, where concepts group into ever larger and denser higher-order associative modules. Recursive relations could lead to new avenues of mathematical analysis to investigate underlying properties in the emergence of semantic complexity, and in particular, how this complexity might systematically shift as a function of varying reinforcements across a range of tasks and experiences across various grounded scenarios.

However, much of computational semantics research (e.g. Rogers & McClelland, 2004) has implicitly advocated for strong *taxonomic hierarchies*. Many computational models are evaluated based on these assumptions from early feature norm studies (e.g. Rosch, 1975). We suggest that there is a parallel between such taxonomic hierarchies and *scale-free* network models, defined by their degree distribution following a *power law*. A fundamental commonality across these networks is the presence of so-called *network hubs*, which have a significantly higher-than-average degree centrality. In our view, these hubs represent superordinate categories like *animal* and *plant*, which have secondary-level leaves, with ever decreasing levels of network degree. Studies on how the *world wide web* is structured strongly indicate a scale-free topological organisation, where a small number of highly influential hubs (e.g. *search engines* and *web services*) provide access to highly recursively structured websites. However, our semantic network topology does not reveal such patterns and instead has *small-world* properties.

The most critical finding of our present study is that alterations in context can dynamically modulate the more lexical network topology, and reshape it based on a more context-relevant structure (hypothesis 5). This context-based-topology is found to be the case across all context comparisons, where the clustering and interconnectivity within semantic

networks in more specific situations reflect the characteristics of a given scenario. Therefore, we argue for a shift away from the assumption of contemporary cognitive science and artificial intelligence that semantic space is constant across different individuals and contexts. For example, Huth, Nishimoto, Vu and Gallant (2012) used 1,705 object and action categories, from a small sample of five male participants, to claim that the semantic space is both continuous and consistent across different people. We suggest that these interpretations are an artefact of the use of linear dimensionality reduction (PCA) and the *same contexts* across all participants. Recently, some theoretical positions (e.g. Hoffman et al., 2018) have started to incorporate the idea that semantic representations might indeed be different across different people. However, our view is more radical than that. Our results support a topology of meaning grounded in real context. Therefore, we advocate both *within* and *between* system-level variability in the topology of cognitive semantics.

Additionally, our results show that not only do different semantic componential dimensions differentially activate concepts but that this can also dynamically shift as a result of changing contexts. Therefore, we argue that the *meaning topology* is multimodal, and our *differential activations quotient* (DAQ) computed across the *neutral-context*, and *context* scenarios can disentangle the relative shifts in importance of underlying semantic components. These results collectively show that dominant dimensions like *vision*, across all concepts, and topic-specific dimensions such as *ingestion*, are weaker for the same sets of concepts when grounded in specific scenarios like *burning house* or *cooking*. On the other hand, weaker dimensions in the neutral context can become more important in specific situations. However, even though some dimensions (e.g. *ingestion*) are more stable across different contexts, these dimensions also undergo dynamic shifts in extreme cases (e.g. *burning house*). Therefore, the relative strength of cognitive dimensions varies across differently grounded contexts (hypothesis 7).

Our adaptation of the *Gini coefficient* and the *Lorenz-curve* show that not all 16 cognitive dimensions are equally important across the semantic network (hypothesis 6). This inequality is captured using *absolute* (number of nodes) and *relative* (Gini coefficient) measures. Regarding absolute measures, *place*, *vision* and *ingestion* dimensions maximally discriminate more than 10% of the concept nodes in the network, while the primitives *surprise*, *upper limb* and *slow*, activate less than 3% of the nodes. Nonetheless, the *upper limb* primitive is more critical across a range of specific contexts associated with movement (e.g. *cooking* and *moving house*). Through our “equality analysis” using the Lorenz-style curve and the *absolute semantic dominance* metric, it is clear that *place* is the single most important dimension, due to its widespread activations across the cognitive semantic network. The semantic network contains associative (e.g. minister-mayor), taxonomic (e.g. season-spring) and functional (e.g. clarinet-saxophone) associations (hypothesis 4) and is not solely structured based on modality-specific delineations (hypothesis 11).

On the other hand, *ingestion* is a less dominant dimension across the entire network, because of its strong preferential activation of food-related concepts, while *slow* only activates 4 concepts across 3/19 clusters, resulting in the lowest levels of absolute and relative semantic importance. The fact that *place* is the most critical cognitive dimension in both absolute and relative measures when it comes to shaping our conceptual networks supports the importance of the real-world, scenes, settings and landmarks in the study of cognitive semantics. This high utility of the cognitive semantic primitive *place* supports our central thesis of the current dissertation. Meaning can and does arise from our everyday surroundings, which is not merely limited to language but also incorporates our real-world physical environment.

Our conceptual topology is strongly dependent on the assumption that the 16 cognitive dimensions are sufficient for explaining much of human cognitive semantics. Given that this is the focus of Binder et al.’s

(2016) neuroimaging-based study, we do not address the validity of the approach per se. Although based on the current results, one might suggest that given the fluidity of semantic topology, it might well be that the 65, more granular, cognitive dimensions initially analysed by Binder et al. might be more suitable. However, for large-scale normative studies like ours, the logistical challenges of including 65 dimensions would perhaps be insurmountable. Instead, now that we provide support for the context-specificity of semantic topologies, future investigations could delimit the set of concepts to a smaller subset, but investigate these across a range of different contexts to evaluate the dynamics of the semantic network. Since our current results have shown the variable nature of cognitive semantic topologies, future studies could test specific hypotheses as to the *types of transformations* one might reasonably predict as a function of manipulating context in subtler ways, in order to understand more gradual transformations approximating real-world semantic tasks.

In our six contexts, the scenarios are discrete. Although this is useful for revealing the non-static nature of semantic spaces, it nonetheless lacks ecological validity. Ideally, in future studies, more nuanced variations in contexts across a small set of fixed concepts might shed light on how *symbol interdependency*, captured by our network topology, shifts from one context to another. We hypothesise that there is a further layer of complexity. In addition to different semantic topologies across contexts, as in our present results, we also expect that within-subject experimental designs might reveal asymmetrical transitions in semantic topology. In other words, people's conceptual topology does not merely shift across different contexts, but the *order of traversing* different contexts itself might also have an impact on the semantic topology. For example, in the hypothetical case of a *neutral context* (*N*) along with *contexts A* and *B*, we predict that there are not merely three different semantic topologies, but *six* different variants. In a transition matrix of 3×3 (set = {*N, A, B*}), where the diagonal indicates no transition, we predict the semantic topology to be identical. However, we

also predict different semantic network topologies when transitioning from a neutral state to a specific context ($N \rightarrow A \neq A \rightarrow N$) as well as in the more realistic form of moving from one context to another ($A \rightarrow B \neq B \rightarrow A$). Therefore, in the case of N , A and B , we predict six distinct topologies. The difference in the topological network reorganisation across contexts will vary as a function of the similarity of contexts, with greater similarity leading to fewer shifts.

6.5.2 Limitations

Next, we turn to some of the limitations of our current study. Although, our results demonstrate that the use of t-SNE can reveal more nuanced semantic associations, the presence of an additional parameter (t-SNE's *perplexity*) is not as parsimonious to simple linear dimensionality reduction methods like PCA or MDS. Moreover, *perplexity* also has a strong influence on the resultant semantic space, which only compounds this limitation. However, we are cautiously optimistic that the benefits of non-linear dimensionality reduction, though increasing model complexity, ultimately will be more successful at capturing semantic interdependencies across different levels of abstraction. Our study is an exploratory starting point for further investigating non-linear dimensionality reduction and its role in cognitive semantic network topography.

Another limitation of the present study is related to the correlation thresholding used for extracting the number of network edges represented in the network's *adjacency matrix*. We use a single threshold across the entire association matrix, which might not be the best way of generating a topology spanning the full concreteness spectrum of a semantic network grounded in a set of basic componential dimensions. We opt for this simple strategy because more sophisticated alternatives in psychological network visualisations are currently still being debated without a consensus (Barfuss et al., 2016; Christensen et al., 2018). However, this might be a limitation for generating semantic network topologies on the basis that

different types of concepts across different contexts might have different *optimal* thresholds. However, given the lack of previous research related to our study, we are unable to determine any *a priori* hypotheses. We hypothesise that more concrete words might need a more conservative threshold (higher than our $r \geq 0.92$) in order to have more associatively and functionally relevant topological properties. However, based on our exploratory network visualisations across different perplexity and correlation threshold, such a high threshold is likely to damage associatively meaningful concept relations for more abstract and intermediary words. Therefore, an open question for future investigations might be to explore a range of suitable thresholding across the concreteness spectrum and cross-validating the results with those obtained independently from neuroimaging studies.

Although not a limitation per se, the network diameter of our semantic network seems to be exceptionally large at 56, especially compared to the random *null distribution* model's diameter of seven. We believe this is associated with our semantic network's concreteness chaining phenomenon, manifested by a gradual transitioning between concrete and abstract words. This chaining by the latent factor of concreteness, not used in the generation of the network itself, is also likely to be the main reason for SemNet's significantly longer diameter. Although we consider this diameter difference to be an exploratory finding, we also consider this exaggerated difference to only apply to neutral networks of the same size.

We believe that our incorporation of the *Gini coefficient* and the *Lorenz-style* curves in analysing semantic networks will be useful for a range of investigations. One might argue that our operationalisation could be improved. Using our incidence matrix, we calculate the Lorenz curves and the Gini coefficient only incorporating incidences of the maximally activated cognitive dimension for a given cluster. This criterion is likely to skew our metrics given that, in some cases, the second, third or even fourth

most dominant dimension might be quite strong contributors to a particular cluster. Including only the most dominant dimension might be overly simplistic, and could be further investigated in subsequent studies.

Another limitation is that spurious relations between concepts are difficult to avoid when concepts are present which do not share much in common with other concepts in the stimuli set. At higher correlation thresholds, “concept islands” emerge, while at lower thresholds, uninterpretable, weak and spurious associations abound. This thresholding dilemma is common across network science and is not a specific limitation of our study. Although we run 5,000 iterations to minimise such spurious associations, recent studies such as Christensen et al. (2018) advocate the use of *information filtering networks* (IFNs), for mitigating this shortcoming. However, this problem is well-known as *sparse structure learning*, where the objective is to reduce dimensionality and complexity in order to increase interpretability while simultaneously ensuring critical information is not lost (Zhao et al., 2011).

Many of the limitations presented above are due to this study being an exploratory investigation even though we did have some explicit hypotheses. We envisage, and hope, that subsequent and more detailed confirmatory research will help overcome some of the shortcomings discussed.

6.5.3 General Implications

Our focus now shifts to the implications of a dynamic semantic network topology and its resultant phenomena, the *relativity of meaning*, to cognitive science and artificial intelligence. This study supports the notion of meaning being dynamic, ever-evolving and prone to alteration as a consequence of our grounded reality. To the best of our knowledge, there is currently no empirical or computational cognitive research investigating this dynamicity of semantic topologies. Our current findings on the

relativity of semantic network topologies might have several implications on cognitive modelling of semantics.

Firstly, most current computational models do not incorporate context, and when included, it is in the form of linguistic topic models (e.g. Hoffman et al., 2018). A natural extension of the work presented in this chapter would be to evaluate ecologically valid grounded models of semantics across a wide range of contexts to investigate corresponding variations in the resultant semantic network topology. Exploring this might help delineate a range of mechanistic accounts that could be tested using empirically derived *ground truths*. Secondly, what types of computational mechanisms could account for local versus global processing of meaning? Our analyses using t-SNE's perplexity measure might be one interesting avenue of future research. Thirdly, how would specific tasks and contexts be incorporated into computational models? Are these merely additional "spoke-type" inputs feeding into a single hub-style representation layer, as in recent models explored by Hoffman et al. (2018)? Would such an approach be even successful beyond smaller toy model implementations? We conjecture that a single hub-representation encompassing various degrees of global to local concepts, across the concreteness spectrum in a dynamic context-sensitive format might be challenging to achieve. In our view, *representational pluralism* might be more suited to address some of these challenges.

Fourthly, and critically, since our semantic topology shifts as a function of context, it is likely that not only associative but also higher-order relational properties might shift. Although our present study does not explore this more specific hypothesis, one could imagine that there are *rules* or perhaps even *laws* that might govern the integration of different primary semantic components. We suggest this because, despite the shifts in the topology across contexts, our results within each context provides reassuringly consistent results. The presence of rules or laws might also explain why even though our meaning space shifts from one scenario to

another, it does so in a reasonably predictable fashion across different individuals. However, this also raises our fifth and final implication for cognitive modelling. To what extent does the topology of semantic networks change across different scenarios *and* specific groups of individuals. For example, in the *burning house* scenario, how would a *firefighter's* semantic space compare with that of ordinary participants? Addressing the above questions will help further our current aims of using cognitive modelling, grounded in the real world, to address complex questions on symbol interdependency through detailed cognitive semantic topography.

Our findings might also have slightly broader implications for cognitive science. In this chapter, we present a network science approach to understanding and mapping meaning. This approach may have advantages for analysing data from cognitive studies using *semantic priming* experimental procedures. Network metrics could be used for detailed comparisons of computational models with behavioural and neuroimaging findings. Another question might be to explore the network structures of *implicit* and *explicit* semantic associations. Our study only explored explicit semantic associations.

However, the respondent instructions for completing the ratings promptly in the present study aims to allow our brain-based CFR task also to capture non-deliberative associations. For example, this includes experimental design aspects like rating only a single dimension per respondent to increase response speeds and the stimuli words automatically changing after selecting a Likert-scale rating (avoiding the extra step of pressing the *next* button). However, we did not collect any response-time data or develop more *Implicit Association Test* type measures (Greenwald, Nosek, & Banaji, 2003), but future investigations could explore this in order to evaluate explicit and implicit semantic network topologies. Lastly, from a cognitive psychological perspective, our neutral context and contextualised topologies suggest that semantic associations between

concepts collected under sterile lab conditions might capture interesting lexical associations such as the concreteness effect, but not be conducive to understanding *real-world* cognitive semantics. We argue a more ecologically valid approach to studying meaning might provide alternative avenues for future research.

We also consider that a dynamic semantic network topology will be particularly important for developing *artificial general intelligence* (AGI) because current state-of-the-art performance in AI is highly domain-specific and non-generalisable. For example, an image classification AI agent trained to classify a range of *common fruits* (e.g. apple, orange, pear) would be mainly performing at chance-levels on the related task of classifying *exotic fruits* (e.g. mango, pineapple, guava). This simple example's limitations stem from recent advances in contemporary supervised learning on millions of human-labelled data points. In our view, this is the equivalent of developing AI with a strong bias for local or specialist meaning encoding, while omitting the long-range associative pathways that lead to the small-world architecture of human meaning.

In AI, there is a well-established literature on robots, both software and hardware instantiations, navigating physical spaces to get from A to B, while engaging in object classification and collision avoidance (see Parhi, 2018, for a review). Nevertheless, almost six decades later, our focus remains to shift toward navigating the more complex and ill-defined “maze of meaning”, that humans traverse on a daily basis. Machine learning and AI “breakthroughs” reported in both the leading academic journals like *Nature* and *Science* as well as by the media predominantly focus on very narrow game playing agents or overly simplistic simulation worlds (e.g. Gibney, 2015). Our work is relevant to the field of AI because it shows that a highly dynamic and context-dependent human conceptual space can be represented more distinctively using non-linear dimensionality reduction and network visualisation. This might help future AIs traverse the complexities of meaning.

6.5.4 Implications for AI

Our penultimate focus, in this discussion, will be to outline some of the pitfalls of AI agents with static semantic spaces and thereby demonstrate the power of incorporating dynamic semantic memories. This argument should also address the more general problem of overcoming the narrow semantic specialisation of most modern AI systems. Let us start by imagining an AI system consisting of circa half-a-dozen ML models integrated into a single deep learning stack¹². Now, let us also imagine that this AI platform contains machine learning models like *word2vec* for semantic representations, a pre-trained deep *convolutional neural network* for vision and also has effectors for interacting in the real world. This AI system goes through tens of millions of training epochs, independently for each machine learning subsystem (e.g. CNN for vision) and all of these models trained together in a stack with interleaved data to ensure the presence of sufficient statistical regularities between language-based and visual inputs. In our analogy, although relevant to a cognitive modelling perspective, we will ignore obvious superficial differences between this system and human development, like the *volume of data*. Now, what can this AI system do? This AI could perform a range of seemingly intelligent tasks with high speed and efficiency. For example, the AI bot could provide useful text-based captions when provided with just an image (e.g. Ramisa et al., 2018). However, how useful would a static semantic space consisting of eight objects (*cup, mobile phone, light camping chair, hamster, wallet, tennis ball, rabbit* and *a wooden stool*) be for the following two situations?

In situation one, the AI bot is asked to *select and place* an object to prop open a heavy door, based on the eight objects it can see. In the second

¹² A concept in commercial data science, popularized by technical consultancies like *Accenture* and *McKinsey Analytics* for integrating or “stacking” multiple independent machine learning algorithms.

situation, the robot is asked to help save the *essential* things from a burning house. We argue that for both these scenarios, our AI agent would fail to make *sensible* choices as a result of the semantic space not being sufficiently dynamic for this task and context. We also note that it would be highly impractical to have to compute *a priori* all possible situations in order to train the AI agent in a supervised manner, as this is unrealistic given the combinatorial explosion of even the more common object use cases let alone *ad hoc* categories such as the ones presented here or more abstract semantic associations. Therefore, AGI especially in the domain of meaning-making, might more realistically rely on a set of primitive semantic dimensions (the *syntax*) and abstract rules (the *grammar*) through which these components are modulated to yield context- and task-specific semantic topologies. In our view, this might help explain both the relative stability and predictability of human meaning-making.

6.5.5 Summary

In closing, although there has been a recent shift towards more “universal semantic theories” (e.g. Vigliocco et al., 2004; Huth et al., 2012; Troche et al., 2014) by highly influential cognitive scientists like Matthew Botvinick (e.g. Pereira et al., 2018), we argue that the basis of this convergence relies on the methodological limitation of not systematically varying context. We do not deny the presence of *more universal* representations of concepts that span a range of contexts, as presented in Pereira and colleagues’ universal decoder of linguistic meaning. We do, however, raise the concern of the usefulness of context-independent semantic spaces because this is rarely the case in real-world scenarios. Our ordinary meaning construal is bounded by the *here and now* forces of not only a given environment but more specifically a *task* and *context*. At times, this meaning space also has to accommodate highly atypical and nuanced *ad hoc categories*, which bear little to no resemblance to previously experienced situations. A semantic topology grounded in a componential

set of basic semantic building blocks helps achieve this and disentangle new meaning in otherwise hidden complexities. It is unlikely that we use a static semantic space for meaning representation in real-world conditions. Our results imply that the human semantic topology varies not only across individuals but also within people in different scenarios. Collectively, these findings support a grounded perspective on cognitive semantics and that there is *no meaning without context*.

Chapter 7

Gender Bias in Grounded Semantics: Network Regularisation and Debiasing

7.1 Abstract

Off-the-shelf machine learning algorithms are increasingly used to make sense of not only language but also our rich visual world. In chapter 4, we demonstrate the feasibility of object co-occurrences representing semantics but subsequently discover limitations of exclusively scene-based meaning in chapter 5. Then, in chapter 6, we revealed the dynamic nature of meaning captured using network analysis grounded in cognitive dimensions. In this chapter, we extend *scene2vec* with an additional “off-the-shelf” algorithm linking visual scenes to linguistic tags. Using this modified *scene2vec* representation, we explore the topical issue of *gender bias* in two popular web-based image repositories - *Google Images* and *Getty Images*. Using network analysis and our novel application of *graphical*

LASSO regularisation to computational semantics, we show that *scene2vec* trained on *Google Images*, but not on *Getty Images*, encodes well-established *gender-occupation* stereotypes from the psychological literature (e.g. *man-doctor* or *woman-nurse*). The presence of context-specific human-like gender biases in *scene2vec* provides a new and scalable method for mechanistically investigating bias. Lastly, we develop a simple debiasing technique, called *semantic feature neutralisation* (SFN), which, in our small-scale semantic model, can selectively target and remove undesirable biases, while leaving desirable gender associations intact.

7.2 Introduction

The term *bias* in cognitive science has several distinct interpretations, depending on whether the focus is on *artificial neural networks*, *decision making* or *belief systems*. In this chapter, we focus specifically on gender biases as a cultural construct. Biases are a disproportionate negative or positive inclination towards something or someone. Human biases are likely to date back to the very roots of prehistoric human development. Modern day culture is expected to determine the ebb and flow of the dominant streams of contemporary prejudices. These *cultural inventions of difference* can be directly associated with clear political and media agendas, but also with more implicit or hidden cultural forces that can be difficult to isolate. An example of the former is an increase in xenophobia in the United States of America following recent anti-immigration rhetoric and also in post-Brexit United Kingdom (Inglehart & Norris, 2016). There is a steady rise in populism in both Western and Eastern societies. Ironically, populism is a global phenomenon (Moffitt, 2016).

A dominant example of implicit biases stems from *gender-occupation* stereotyping. Over the last three decades, there has been a steady decline in women seeking University-level computer science qualifications. Fuelling

this state of affairs are popular and unhelpful cultural stereotypes of “geeky unpopular guys” who are obsessed with technology (see Mercier, Barron, & O’Connor, 2006, for a detailed discussion). These biases can act as barriers to diversifying traditional STEM disciplines (Cheryan, Master, & Meltzoff, 2015). Relatedly, a widely replicated social psychology example of gender-occupation biases associates MEN with SCIENCES and WOMEN with LIBERAL ARTS (e.g. Bennett, 1982; Greenwald, McGhee, & Schwartz, 1998; Rudman & Kilianski, 2000; Greenwald et al., 2009).

Unfortunately, although some psychologists can unearth human biases, others perpetuate them by crafting convincing narratives which happen to neatly fit the pre-existing belief systems of the audience. For example, some theories (e.g. Alexander, 2003) claim that women prefer the colour *pink* while men prefer the colour *blue* from evolutionary differences (men hunting for animals with the *blue* sky in the background and women picking *reddish* berries). Kandola and Kandola (2013), the occupational psychologists specialising in gender differences, point out that evolutionary psychological theories like those of Alexander (2003) are a “back-projection of later gender divisions on to earlier ways of life” (p.17). These biases are also justified and amplified by the popular press.

Understanding the cognitive semantics underlying biases can have tangible benefits for society. Throughout the present thesis, we repeatedly emphasise the importance of real-world grounding of semantic associations encoded in computational models. Our central thesis encourages avoiding arbitrary hand-coded features, which, in isolation can lead to a vicious cycle of encoding and decoding the meaning representations one determines *a priori*. Additionally, we advocate for grounded knowledge representations to circumvent the circularity of the symbol grounding merry-go-round. In this chapter, we evaluate whether our grounded scene-based representations are ecologically valid. If scene-based grounding is psychologically meaningful, then it is only natural also to expect that biases or stereotypes commonly encountered in humans

should also be present in our computational models of semantic cognition. Therefore, exploring human biases provides a litmus test of the psychological relevance of our scene-based semantic representations and holds the potential to investigate prejudices in semantic cognition mechanistically.

7.2.1 Historical and Empirical Foundations of Bias

In social psychology, the main impetus for a rigorous investigation of bias dates back to work distinguishing *implicit* and *explicit cognition* (Schacter, Bowers, & Booker, 1989). Biases frequently reside in the realms of implicit or latent unconscious associations. Explicit cognition, on the other hand, is related to higher levels of *controllability*, *intentionality* and *awareness* (Nosek, 2007). Our focus in this brief overview of bias will be limited to the implicit cognition literature, focusing on the widely documented *Implicit Association Test* (IAT).

Greenwald and Banaji (1995) define implicit bias as previous experiences shaping current or future performance even though the earlier prior experiences are not directly recalled and therefore are unavailable to introspection (p. 5). The IAT operationalises the psychological measurement of implicit biases and evaluates the associative strength between a given set of concepts. Although numerous variants exist, the most widely used IAT (Greenwald, McGhee, & Schwartz, 1998) measures exemplars from four concepts (e.g. MALE/FEMALE and SCIENCE/LIBERAL ARTS) using only two response options associated with two concept pairs in a given experimental block. The primary dependent variables of interest are *response latencies* and *error rates*. When participants are asked to compare similar concepts, response latencies and error rates are lower than when asked to compare very different concepts. Since these measures are dependent on reaction times in milliseconds, one generally accepts the IAT as a procedure for extracting implicit associations. In Nosek and colleagues' *gender-science* IAT, where the

concept categories were MALE and FEMALE (e.g. *he* and *she*) and SCIENCE and LIBERAL ARTS (e.g. *physics* and *literature*), the results supported a robust “automatic association of male with science and female with liberal arts compared with the complimentary pairings” (p. 108).

Greenwald et al. (1998) used a wide range of phenomena to demonstrate the validity of IAT, which consisted of both innocuous demonstrations of biases (e.g. *flowers* are more pleasant than *insects*) as well as more disquieting effects of strong racial biases when performing the test with images of WHITE and BLACK individuals and PLEASANT and UNPLEASANT word associations.

7.2.2 Implicit Gender Biases

Nosek, Banaji and Greenwald (2002) conducted a large-scale online study with over 600,000 tasks across a range of implicit and explicit evaluations spanning *race*, *age* and *gender*. In the *gender-career* IAT version, the four concept groups were MALE and FEMALE (e.g. *boy* and *girl*) along with CAREER and FAMILY terms (e.g. *executive* and *children*). The results showed a strong and robust preferential implicit association between *male* and *career* and *female* and *family*.

At a societal level, gender biases constitute a significant issue for women deciding to study mathematics or science or pursue technical careers in, for example, *data science* - an applied field at the intersection of statistics and computer science. Such leanings are particularly surprising when considering that on standardised high-school level mathematics tests, girls and boys perform similarly and are equally prepared for studying STEM subjects (science, technology, engineering and mathematics). Nonetheless, at college-level, men far outnumber women in most STEM subjects (Zafar, 2013). In a relatively recent study, Reuben, Sapienza and Zingales (2014) showed that in a hiring scenario, when the only information available to hiring managers is appearance (revealing gender), men are twice as likely to be hired for a mathematical task than women.

Interestingly, the scores of a Gender-Occupation IAT can account for this difference. Therefore, gender biases directly impact hiring prospects.

Gender biases along with the negative implications of hiring prospects for women in a numerical capacity indicates a strong need to investigate the underlying semantic associations of gender bias. Finally, given McKinsey's prediction of artificial intelligence permeating all facets of our lives (Pyle & San Jose, 2015), we suggest it is also essential to understand biases in machine learning.

7.2.3 Biases in Machine Learning

People intuitively think that artificial intelligence (AI), and machine learning (ML) do not suffer from human-like biases. A naïve, but common, assumption of the general public is that AI/ML is "just mathematics". What most people outside of the niche fields of AI, ML and cognitive modelling are not aware of is that the vast majority of thriving computational intelligence applications stem from massive troves of unstructured and structured data. A wide range of different algorithms exploit this data using predominantly unsupervised classification algorithms. In this thesis, we have consistently favoured real-world or grounded informational inputs for our cognitive computational models. However, in doing so, although there are undoubtedly many advantages, we have potentially discounted one critical shortcoming - the introduction and perpetuation of harmful human biases to cognitive modelling. To foreshadow our subsequent discussions, if the inclusion of such biases were feasible into computational cognitive models, then this might have advantages and disadvantages. However, we first need to establish whether or not human-like biases exist in our current iteration of *scene2vec*.

It has been well-known for several years that automated statistical learning systems are strongly prone to so-called *machine prejudice*. One of the first instances of a well-documented and widely publicised case stems from the domain of online advertising and racial discrimination. Sweeney

(2013) collated a list of 2,000 names, which were particularly suggestive of race. For example, the names *Trevon* and *Lakisha* are more common for black individuals, while the names *Brendan* and *Katie* for white people. Sweeney entered these names on *Google.com* and *Reuters.com* and recorded the types of ads that were automatically generated based on the name entered. On both websites, entering “black names” led to significantly more arrest-related ads than when entering “white names”. This unfairness can potentially have significant negative consequences, consciously or unconsciously, if, for example, prospective employers perform Google searches for job applicants’ names. In this case, the bias was a direct consequence of automated click-through systems that map the statistical regularities of words with specific online ads. Similarly, there are other domains where machine learning biases have crept in, such as *credit scoring* (Hardt, Price, & Srebro, 2016), *news* (Ross & Carter, 2011) and, unfortunately, even *criminal sentencing* (Angwin et al., 2016). In all cases, there is a disproportionate impact on minority ethnic groups and women.

In a recent computational experiment, published in *Science*, Caliskan, Bryson and Narayanan (2017) explored the presence and effect size of biases in semantics derived from an “off-the-shelf” distributed language-based model. Caliskan et al. used the well-known *GloVe* (Global vectors for word representation) word embeddings (Pennington, Socher, & Manning, 2014). *GloVe* is a state-of-the-art word embedding algorithm for dimensionality reduction and amplification of surface-level language co-occurrence statistics. These embeddings have specific semantic properties that make them particularly appealing. Collobert et al. (2011) have shown that words like *France* are closer to *Italy* and *Austria* using embeddings because vectors represent concepts, and subtraction/addition of these vectors is meaningful. For instance, differences in vectors *England* and *London* is similar to the same subtraction operation applied to *France* and *Paris* or *Austria* and *Vienna*.

Similarly, by subtracting *man* from *woman*, gender differences are derived with the resulting vector being almost parallel to the vector following the subtraction of *king* from *queen* (*king:man :: woman:queen*). Caliskan and colleagues used GloVe, with 300 dimensions, pre-trained on the most extensive web corpora, called *Common Crawl*, consisting of 840 billion case-sensitive tokens, with 2.2 million unique word tokens. Caliskan et al. developed an analogous computational variant of IAT, called WEAT (Word-Embedding Association Test), which uses the 300-dimensional vector representations for a given set of target words to calculate the *cosine similarity score* between all words of interest. Using this technique, they successfully replicated all biases documented in Greenwald et al. (1998). This ranged from the relatively innocuous biases of *flowers are more pleasant than insects* and *musical instruments are more pleasant than weapons*, to the more offensive racial and gender biases of, respectively, European American names being more pleasant than African American names and male names being more associated with career words, while female names being more associated with family words.

Sample imbalances can contribute to these biases in language-based models - where one or more classes of words or objects are either over- or under-represented in different contexts (e.g. documents). In the case of classifier algorithms, when a class is overrepresented, then most algorithms with the objective function of achieving the highest accuracy level will inadvertently be superior at predicting the majority label. The cost of misclassification for the majority label is too high, unlike the case for minority labels, which have lower impacts on a given algorithm's overall accuracy level (Zhao et al., 2017, p.2). These findings from the literature on bias/prejudice and machine learning also support the idea that human biases are indeed statistical regularities with sufficiently large signal-to-noise ratio such that they are successfully captured in language-based vector-space models of semantics, a core paradigm of much of contemporary ML and AI research.

7.2.4 Gender Biases in Machine Learning

Jenset and McGillivray (2017) conducted a detailed investigation on, once again, word embeddings, but over a period of 100 years to investigate the temporal dynamics of biases. Jenset and McGillivray used Google Books and the Corpus of Historical American English to create nine distinct embeddings, each trained on a decade from the 1900s onwards. These temporally-distinct word embeddings were used to identify a phase-shift, between the 1960s and 1970s. This shift is due to a relative increase in adjectives such as intelligent and logical co-occurring more with women post-1960s. Another underlying factor determining this shift was the predominant use of appearance-based adjectives used for women pre-1960s (e.g. attractive or fashionable). This research highlights the biased nature of language and uses chronologically ordered embeddings to quantify the impact of the women's movement objectively. Jenset and McGillivray's results provide empirical support for time-dependent statistical regularities of language shaping human biases. Unsurprisingly, ML models trained on this biased data will display gender-biased semantic associations.

7.2.5 Awareness of Debiasing Machine Learning

Successfully debiasing machine learning is critical for addressing the needs of a fair and equal society. An early precursor to current worries about algorithmic fairness emerged from the framework *Discrimination Aware Data Mining* (DADM), in the context of association rule mining. Pedreshi, Ruggieri and Turini's (2008) far-sighted DADM philosophy, based on the *civil rights law*, states "discrimination refers to [the] unfair or unequal treatment of people based on membership to a category or a minority, without regard to individual merit" (p. 1). DADM is quantified using a term they coined, *α -protection*, which measures the discriminatory power of a particular association rule used for database targeting. More

recently, since 2014, a community of commercial and academic researchers in *machine fairness* have initiated an annual conference on Fairness, Accountability, Transparency Machine Learning (FAT ML)¹³. These initiatives help identify and generate greater awareness of the underlying problem and find solutions for debiasing algorithms. Nonetheless, the issue of biases in algorithms has largely gone unnoticed in the cognitive modelling community.

7.3 Investigating Gender Bias in Grounded Semantics

7.3.1 Objective

Previous machine learning research (Caliskan et al., 2017) has demonstrated that gender biases are prevalent in word embeddings trained on large-scale language corpora. In our present computational study, we seek to investigate whether or not similar prejudices are also commonplace in our scene-based representations (scene2vec). We aim to contribute to the currently limited but fast-growing extant literature on machine learning fairness, in four distinct ways. First, we want to investigate whether scene2vec contains gender biases. Second, we want to develop a bias metric tailored explicitly for evaluating the bias within semantic association networks and not a 2- or 3-dimensional Euclidean embedding space. Third, if scene2vec contains gender biases, we want to understand whether or not shifting our data source for acquiring the photographs from Google Images to Getty Images reduces this bias, given that Getty Images has resilient human editorial controls, unlike Google Images. Fourth, if scene2vec representations contain gender biases, we intend to evaluate a scalable and efficient debiasing approach by applying a so-called late stage adjustment, by identifying the biased dimensions and selectively extracting this for

¹³ <https://www.fatml.org>

inappropriate concepts. In our view, this debiasing technique has the advantage of neither requiring re-weighting of raw data (early stage) nor adjusting the objective function of the learning algorithm (middle stage).

7.3.2 Methods

7.3.2.1 Stimuli for Computational Models

In this computational study, we use a set of 60 concepts (see *table 7.1*), split into ten categories of six concepts. The main *gender* category consists of three male and female words to minimise noise-based variations in gender associations. Across all quantitative comparisons of gender biases, we use averaged *male* and *female* concepts. We have two concept categories, which we predict to reveal inappropriate or societally undesirable gender biases, which are *business* and *occupation*. The *clothing* category is included to reveal strong but “appropriate gender biases”. All other concept categories are predicted to be gender-neutral. In *table 7.1* all 60 concepts are shown, for each of which we download ten naturalistic-looking photographs from both *Google Images* and *Getty Images*. As is the case for chapters 4 and 5, in this experiment we also manually exclude non-naturalistic images such as marketing logos, cartoons and diagrams.

Person / Gender		
man	male	gentleman
woman	female	lady
Animals	Business	Clothing
cat	business	trousers
dog	finance	shirt
rabbit	meeting	skirt
duck	deal	dress
chicken	trade	suit
tiger	marketing	coat
Fruit	Home	Nature
apple	bathroom	river
orange	kitchen	forest
lemon	bedroom	mountain
banana	garden	jungle
peach	hallway	sea
lychee	TV	lake
Occupation	Shopping	Vehicles
professor	grocery	car
CEO	supermarket	truck
doctor	shopping	bus
nurse	purchase	train
teacher	bakery	limousine
assistant	sale	van

Table 7.1: The 60 concepts used in the gender bias experiments. The concepts are grouped into ten categories.

7.3.2.2 Computational Modelling

As the foundation of our present modelling simulations, we use the identical methodology used in the *scene2vec* modelling of chapter 5, which includes both the object co-occurrence cues from Zhao et al.’s (2017) scene parsing network called *PSPNet*, as well as Microsoft’s Emotion API-based facial expressions classifier. Also, like in our previous *scene2vec* implementation, we once more use the same neural network architecture and parameters during concept training. However, from the limitations discovered in Chapter 5, namely that a *limited set* of vision-only semantic representations constrain the encoding of more abstract concepts sufficiently well, we extend *scene2vec* one final time in our dissertation. Given that Zhao et al.’s network uses the standard 1,000 ImageNet classes, we expand our *scene2vec* representation significantly using Microsoft’s

Azure's Vision API's auto-tagging algorithm¹⁴, which has 10,000+ object classes. Once again, we use a pre-trained, widely-accessible "off-the-shelf" deep learning algorithm, which is a widespread practice in recent ML publications in this domain (e.g. Caliskan et al., 2017) because it aids the reproducibility of computational experiments.

As we initially discuss in chapters 3 and 4, from the computational models of Goldstone and Rogosky (2002) and Louwerse's (2007, 2011) *symbol interdependency hypothesis*, we believe in the mutually reinforcing property of perceptually grounded and linguistic information. Therefore, we incorporate Microsoft's auto-tagging algorithm to include additional language-based features. However, these features are merely a "proxy" for expanding the limited 1,000 ImageNet classes. Therefore, our raw stimuli are still only naturalistic photographs, but our scene2vec features now incorporate linguistic cues.

Auto-tagging algorithms (e.g. Wang et al., 2006) have been useful for coping with the increase in unstructured online data and have been especially important in the field of *information retrieval* (IR). Auto-tagging algorithms are used for encoding and storing a wide range of interrelated visual associations between different images in conjunction with labels or tags. For example, Wang et al. (2006) used a collection of 2.4 million web-images, embedded in a vast and diverse array of naturally occurring texts and images to predict suitable tags for new images not used during model training.

Until now in our thesis, all the scene-based computational models (excluding the language-based comparisons) consist of object-level features (vectors encoding whether or not a particular object or emotion is present). However, with the inclusion of image-tagging, we also incorporate language based associations with the visual world. Although the model input is still consistent with previous chapters (naturalistic photographs),

¹⁴ <https://docs.microsoft.com/en-us/azure/#pivot=products&panel=ai>

we are now utilising the full potential of symbol interdependency - both grounded and symbolic associations. In *figure 7.1* we display a small set of example tags generated for four specific photographs.

In the four images shown in *figure 7.1*, we notice that the tags generated are very poor at capturing object co-occurrences, which is why we interpret our auto-tagging addition as a higher-order linguistic/associative property of the *scene2vec* representation. This lack of object-level information suggests minimal overlap between these tags generated and the object-level information encoded from the PSPNet output. However, when there is duplication, the “duplicated feature” is omitted. The algorithm first represents PSPNet features, and then emotional information is included in a second stage, and the final stage incorporates the tags, with no duplication of cues.

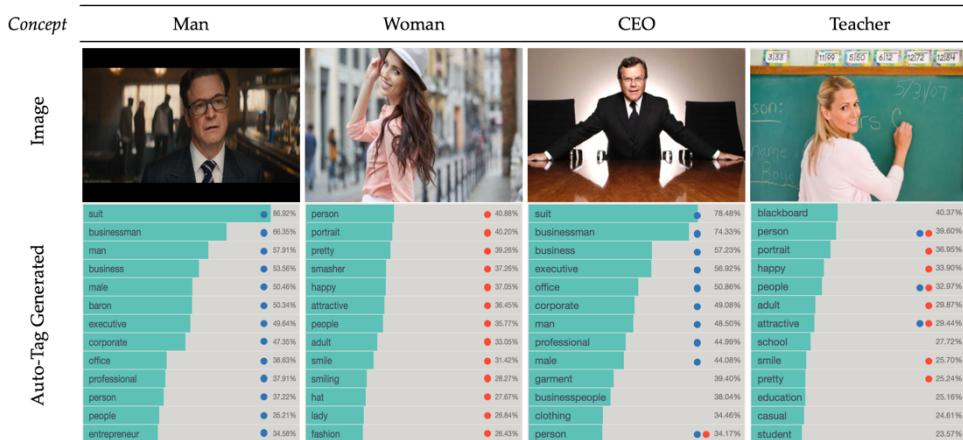


Figure 7.1: Depiction of four concepts, along with an example image and automatically generated tags (only the top 13 tags shown, but the analyses include up to 100 tags per image). We show the tags' confidence scores (%) and have manually placed blue (male) and red (female) dots adjacent to each tag from co-occurrence in the male or female examples shown.

7.3.2.3 Semantic Network Regularisation

In chapter 6 we empirically advance the idea of using semantic network models not only for visualisation purposes but also for investigating the topological properties of the resultant semantic network. However, we also outline that one of the shortcomings of that approach is

that the modeller has to determine an arbitrary cut-off threshold for visualising only the strongest correlations. This is not in keeping with our central thesis - grounding semantics in the real-world with little or no hand-coded or arbitrary modeller-determined rules. Nonetheless, it is not feasible to preserve all correlation coefficients, as the subsequent network visualisation becomes a densely interconnected “ball”, which nullifies interpretability. On the other hand, if a very high correlation threshold is selected, most nodes end up forming sparse islands with a small handful of nodes connected to one or two other nodes, which does not aid network interpretability. Therefore, the most suitable correlation cut-off threshold is somewhere between these two extremes, which is both visually interpretable and sufficiently parsimonious and rigorous to yield consistent results.

One of the main shortcomings of network visualisation in psychology, especially psychometrics and clinical psychology, is the replication of networks structures in direct and conceptual replications (e.g. Forbes et al., 2017). Therefore, in addition to addressing the research questions on gender bias in scene-based representations, a secondary objective of the research in this chapter is to offer a more rigorous approach to representing cognitive semantic networks, given our earlier arguments for the advantages of using network models. We aim to increase parsimony and interpretability of network techniques applied to cognitive semantic models, by *estimating* networks based on finding an acceptable equilibrium between *false positives* and *false negatives*, minimising the probability of spurious network edges.

In a recent tutorial aimed at experimental psychologists interested in using network modelling for measuring the structure of psychological constructs, Epskamp and Fried (2018) outline some of the core shortcomings of network analysis applied to psychological data. In this section, we briefly provide an overview of their core arguments, before our

novel application of their guidelines to our modelling of semantic networks.

In the domain of psychometrics, according to Epskamp and Fried (2018), network analysis has seen a great deal of prominence in recent years (e.g. McNally et al., 2015), with many experimental psychologists (e.g. Schmittmann, et al., 2013) using it to replace *latent variable modelling*. One method for dealing with spurious links/edges in network models is the use of regularisation to estimate *partial correlation matrices*. Given that the correlation matrix determines the adjacency matrix, which feeds into the network model, it is the correlation matrix that needs to be reliably represented. In regularisation methods, common in the machine learning field, the aim is to estimate a statistical model by penalising model complexity. In the case of our specific network consisting of 60 concepts, *any* simple correlation matrix will be at the *upper boundary* for complexity given that even negligible correlation coefficients near zero are still represented as a parameter in the dense correlation matrix. Regularisation methods, however, ensure that relationships that are unlikely to be sufficiently statistically reliable are excluded from the partial correlations, leading to a *sparse model*, in which only a small subset of associations are represented (Epskamp & Fried, 2018). It is worth noting, however, that regularisation does not lead to the omission of specific variables (in our computational model, these are the 60 concepts), as suggested by Christensen et al. (2018), which has garnered some groundless criticism of regularisation techniques in network modelling in psychological model development. We briefly outline the criticism, followed by a rebuttal, which then is followed by the details on network regularisation.

Christensen and colleagues (2018) claim that “[the] shrinkage of correlations below a certain threshold [when using regularisation] also contributes to reduced reproducibility because variables can be eliminated based on statistical significance rather than theory” (p. 11). Clearly, in our case, the “theory aspect” is not essential per se, since we are modelling

concepts of interest as opposed to hypothesised psychological constructs representing distinct processes. Nonetheless, Christensen et al.'s interpretation of what statistical regularisation does to a network model is inaccurate. Unlike step-wise regression, for example, which is an inferential model selection method, and does indeed omit variables, regularisation in network analysis removes spurious edges and not variables. Therefore, in our case, regularisation will not lead to a smaller set of concepts being represented, but a sparser model with fewer links between concepts, as these associations/edges are minimised (Epskamp & Fried, 2018).

A *partial correlation matrix* between n concepts represents the interrelationships present between concepts on a pair-wise basis after statistically controlling for all other associations. In the case of network modelling, when a partial correlation coefficient between the two concepts is precisely zero, no edge is drawn between the respective concepts. However, Epskamp and Fried (2018) highlight some problems with exclusively using partial correlation matrices to determine the "*ground truth*" network structure. Costantini et al. (2015) also highlights this problem of spurious edges and relates this to *false positives*. Sampling variations in the underlying network data, in our case, the raw signals feeding into the *scene2vec* representation, also means that even when two concepts are conditionally independent (e.g. no overlap in grounded features), the partial correlation estimates are likely never to be exactly zero. Epskamp and Fried (2018) discuss one approach that has seen a great deal of popularity, not only in network modelling but partial correlation analysis more broadly, which consists of running statistical significance tests on all pair-wise partial correlations estimated. Failure for a partial correlation coefficient to reach statistical significance then lead to that association not being included in the network. Unfortunately, Epskamp and Fried highlight the main shortcoming of this approach - the *problem of multiple testing*, which can invalidate the original significance threshold set. Moreover, correcting for multiple comparisons is not entirely suitable

either in network models, because of the accompanying loss in *statistical power* (Costantini et al., 2015).

Currently, *LASSO regularisation* is the state-of-the-art approach for overcoming the limitations above while ensuring both interpretability (“genuine sparsity”) and replicability. Although LASSO regularisation (*least absolute shrinkage and selection operator*) is a widespread regularisation technique and is also used in regression (e.g. Hans, 2009), it is particularly useful in the domain of network modelling because of LASSO regularisation’s ability to produce not only partial correlation estimates near zero but estimates that are *precisely zero*. This is achieved by *limiting* the sum of all absolute partial correlation coefficients, which has the effect of *shrinking* all correlation coefficients, leading to some becoming precisely zero (Epskamp & Fried, 2018, p. 5). LASSO also uses a tuning parameter, λ (lambda) which determines the level of sparsity. In cases where λ is set to a low value, only a few edges are omitted, which leads to most spurious edges remaining, while at the highest level (relative to the highest absolute partial correlation coefficient) most edges are omitted, which leads to a highly sparse network, and an increased likelihood of *false negatives*. Then one can select a range of λ tuning parameters and visualise a range of different network models, from a fully connected to an entirely sparse network structure.

Deciding on a suitable λ value seems to be reminiscent of our correlation thresholding dilemma. However, instead of subjectively selecting “a suitable” network structure, the best network in LASSO regularisation can be selected by minimising some *objective information criterion*. Following Epskamp and Fried’s (2018) guidelines for psychological network models, we minimise the *Extended Bayesian Information Criterion* (EBIC). According to Epskamp and Fried, the EBIC has been shown to reveal the real network structures of systems where one expects to find a sparser network topology (Barber & Drton, 2015). In all our network models, we use the default hyperparameter for EBIC = 0.5, as

suggested by Foygel and Drton (2010), which is a more conservative benchmark that aims to reveal sparse networks of higher specificity, with a slightly elevated risk of false negatives. Given our objective of investigating gender biases in grounded semantic networks, we opt for this more parsimonious threshold. We use this regularisation approach to reveal the most robust network edges. This LASSO-EBIC derived network is then used to identify concepts connected *directly* (path length of one) with male {*man*, *male*, *gentleman*} and female {*woman*, *female*, *lady*} concept nodes. We use three different gendered concepts to represent both genders to increase the robustness of gender bias results as relying on a single gender concept might be noisy. Even though the network distances between all 54 concepts (60 - 6 gender-specific concepts) and *male* and *female* concepts are computed, only the concepts directly associated with the gendered concepts are evaluated for gender bias, as these are the concepts with the highest proportion of bias.

We are the first to apply this *graphical Lasso* with *EBIC optimisation* to the domain of cognitive semantic networks, both empirical or computational. In doing so, we would like to set a new precedent of not arbitrarily and qualitatively determining correlation thresholds which subjectively satisfy specific modeller-conceived properties. In summary, we believe that this additional complexity in our network modelling method can be counterbalanced by revealing meaningful semantic network interrelationships and avoiding the over-interpretation of unreliable network topologies. This more rigours approach is particularly important for the present chapter, which seeks to investigate the psychological phenomenon of *gender bias*.

7.3.3 Results

7.3.3.1 Correlation of Google Images

In our study, the semantic regularities grounded in scene-based naturalistic photographs is operationalised using our extended *scene2vec* model. In our first study, we visualise the hidden units of the *scene2vec* neural network trained exclusively on *Google Images*, using a network plot (see *figure 7.2*), with a modeller-defined correlation threshold of $r \geq 0.50$. This threshold is set by generating a network with a single component (all concepts are linked together), in order to facilitate interpretability.

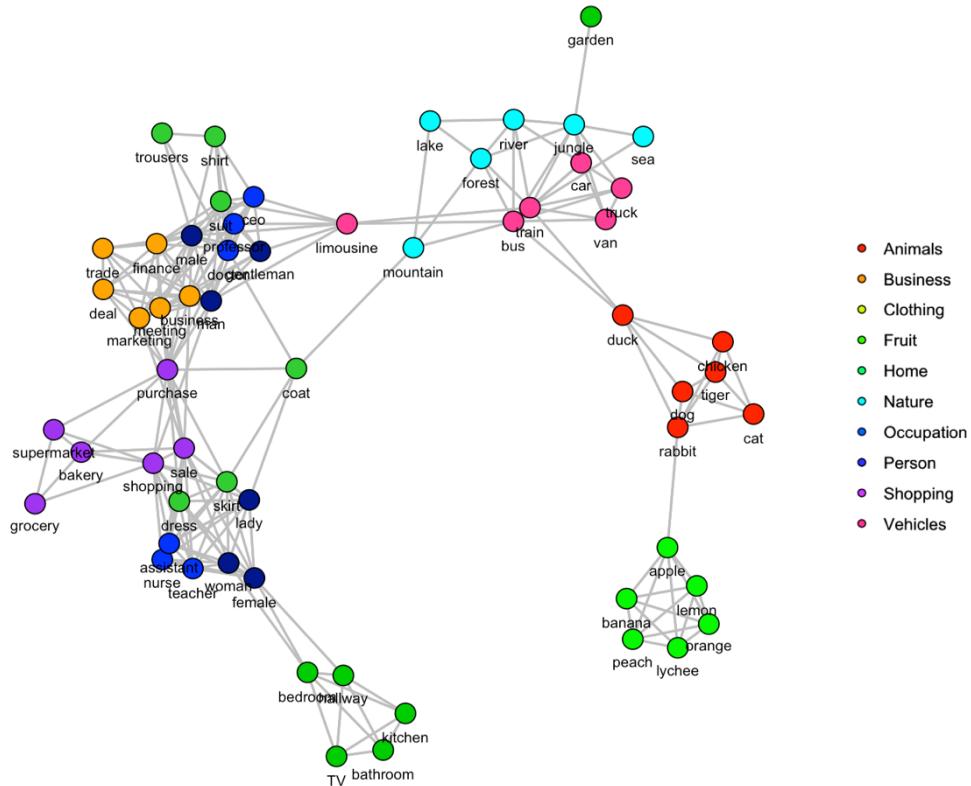


Figure 7.2: Visualisation of the semantic networks grounded in Google Images. Links are only depicted based on the manually-set correlation threshold of $r \geq 0.50$.

In *figure 7.2* we see that the vast majority of the 60 concepts we are modelling using *scene2vec* grounded in *Google Images* are grouped by their semantic categories such as *fruit*, *animals*, *vehicles*, *nature*, *business* and *shopping*. This is also the case for the category *home*, except for the concept

garden, which, somewhat understandably, is more closely associated with the *nature* category, and directly linked with *jungle*. Thus, seven out of ten categories are grouped by their category or a highly related surrogate group. However, this is not the case for the concepts in the category *person* {*man, male, gentleman, woman, female, lady*}, *clothing* {*trousers, shirt, skirt, dress, suit, coat*} and *occupation* {*professor, CEO, doctor, nurse, teacher, assistant*}.

First, the *person* concepts are split neatly into two distinct network regions, without any edges between *male* and *female* concepts. Second, the *clothing* concepts are split according to gender boundaries, such that *trousers, shirt* and *suit* are grouped with *male* concepts while concepts *skirt* and *dress* are grouped with *female* concepts.

Interestingly, the word *coat* is a *central node* in this network, connecting *trousers, skirt, purchase, lady* and *mountain*, therefore spanning the gender divide and also meaningfully associated with concepts from different categories. However, we believe, that this by itself does not necessarily determine gender bias per se, considering that these clothing items are predominantly gender-specific¹⁵. In the case of the only unisex item (the *coat*), we find that this concept is positioned in-between the two gender extremes, which lends qualitative support to the meaningfulness of the network extracted. Third, the concepts within the *profession* category are split into two distinct network groups, corresponding to a sharp gender divide. The professions *professor, CEO* and *doctor* are tightly grouped with the *male* concepts, whereas the professions *nurse, assistant* and *teacher* cohabit the network region with *female* concepts. Fourth, and quite strikingly, the *business* category, although clustered together, is exclusively in the “male region” of the network with direct links to the *man, male* and *gentleman*. This network topology of *scen2vec* based on photographs semi-automatically extracted from *Google Images* suggests an extreme gender bias

¹⁵ The notion of gender fluidity might contradict this interpretation, which we discuss in section 7.4.

wherein men are associated with more powerful and lucrative jobs while women with more care-related and subservient positions.

7.3.3.2 Correlation of Getty Images

In our second network topology (see *figure 7.3*), we visualise the hidden node representations of the statistical regularities of *scene2vec*, but this time, grounded in *Getty Images*. This second image source consists of a professionally edited (left-leaning editorial stance) database of photographs curated for a wide range of private and public purposes. At first glance, there are many surface similarities between this network topology and that grounded in *Google Images*. Like in the *Google Images* network, this network is also broadly structured around concept categories. In specific instances where this is violated, the differences are still semantically meaningful. For example, towards the top of the *Getty Images* network, the concept *sale* is grouped with the *business* concepts instead of the *shopping* concepts {*shopping, purchase, supermarket, grocery, bakery*}.

Similarly, like before, the concept *garden* is grouped with *jungle*, and *forest* in the *Getty Images* network. Lastly, the concepts *chicken* and *duck*, are grouped with *shopping* concepts like *purchase* and *bakery* as well as the *fruit* cluster. *Clothes* are grouped in gender-specific network territories, although, in this network, concept *coat* is in the *male* cluster. However, most interestingly, as predicted, when *scene2vec* is grounded in *Getty Images*, gender bias is not as prevalent as is suggested by the more gender-neutral grouping of the *profession* and *business* concepts. Except for the concept *assistant*, none of the five other professions is directly linked with a particular gender concept. However, subtler gender biases do seem to prevail. Although this was not predicted, there seems to be a linear continuum of professions that emerges, and from left-to-right, moving from more stereotypically feminine gender roles of caring to more male roles of power (*assistant* → *nurse* → *teacher* → *doctor* → *professor* → *CEO*).

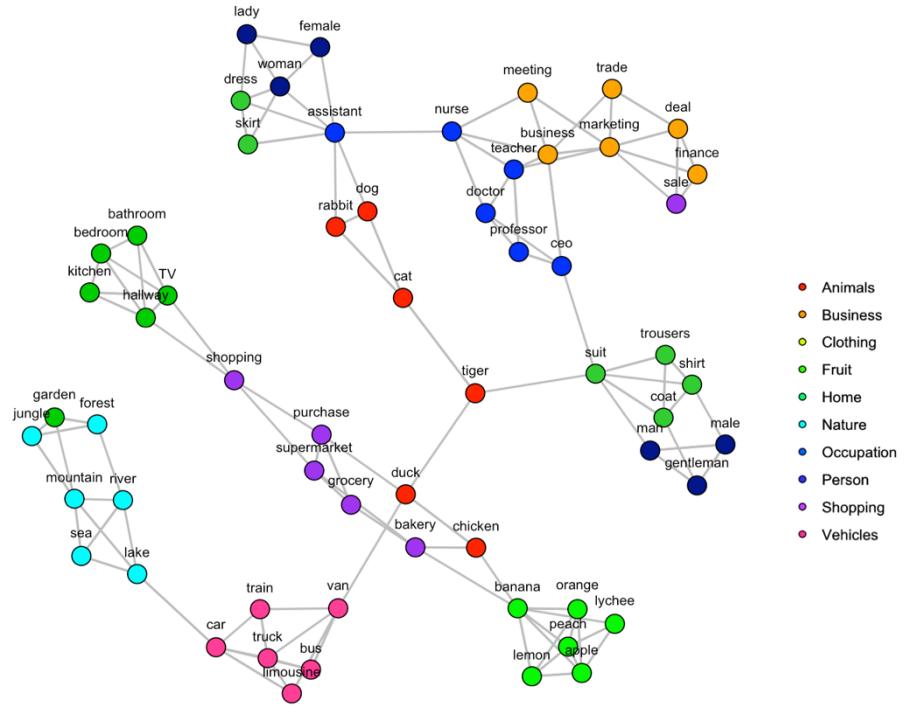


Figure 7.3: Visualisation of the semantic networks grounded in Getty Images. Links are only depicted based on the manually-set correlation threshold of $r \geq 0.50$.

7.3.3.3 Network Centrality

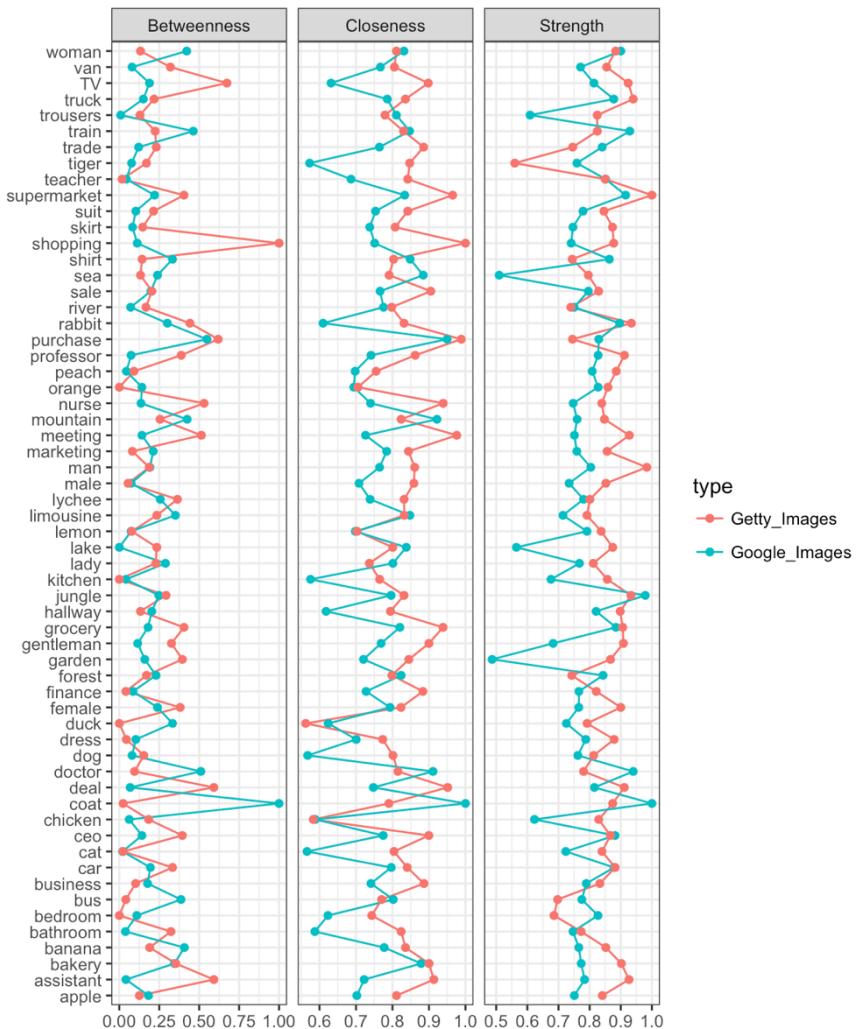


Figure 7.4: Three different network centrality measures (Betweenness, Closeness and Strength) across the Getty Images and Google Images networks.

All network centrality measures are normalised to allow for comparisons across the different topologies of the networks grounded in either *Google Images* or *Getty Images*. The network centrality measures, figure 7.4, show some variability across both networks. For example, the *betweenness* centrality metric indicates that in the *Getty Images* network *shopping* is highly interconnected, while in the *Google Images* network, this is the case for the concept *coat*. However, these three network centrality measures are probably most useful for suggesting that all concepts are relatively equal within their respective networks. However, the *Getty Images* network displays greater *closeness* and *strength* centrality, because

the concepts are more distributed unlike in the case of the Google Images network. However, because of the relatively small size of the overall network, we do not generalise any of these exploratory observations of network centrality to computationally-derived semantic networks more broadly.

7.3.3.4 Regularised Network Visualisation

Next, we re-run our network visualisations using *graphical LASSO-EBIC* regularisation. As argued above this is expected to increase the reliability and reproducibility of our interpretations of the network topologies and reveal whether or not the gender bias for occupations observed in our unregularised networks also manifests itself in the regularised networks. In *figure 7.5A* and *B*, we can see that the regularised networks have a sparser set of strong network edges and a denser proportion of weak (some spurious) links.

In our network grounded in photographs from *Google Images* (*figure 7.5A*), as was the case for the unregularised network, in this regularised version, the concepts are predominantly grouped by the concept categories. Once more, when concepts deviate from this pattern it is typically still associated in a semantically meaningful manner (e.g. *garden* with *jungle*). Although, in the regularised network based on photographs from *Getty Images* (*figure 7.5B*), the concepts *duck* and *chicken* form their own distinct “island” with no dominant links to other concepts. In the unregularised network, these two concepts are still connected with other animals as well as food items. However, the regularised *Getty Images* network suggests that *chicken* and *duck* are more closely related to *just* each other, which is reasonable given that *Getty Images* contains more professional photographs of meals for these two concepts, which correspond to overlapping objects (e.g. *plate* and *cutlery*) and tags (e.g. *dining* and *meal*).

Most importantly, the regularised *Google Images* network lends further support to the unregularised network’s gender-biased structure,

because here we once again see two distinct regions of the network, whereby *male* and *female* concepts are respectively grouped stereotypically either with the professions *professor*, *CEO* and *doctor* or with roles *nurse*, *teacher* and *assistant*. These groupings have been highlighted using dashed blue and red lines, respectively corresponding to the male and female network regions. These areas are outlined manually merely as a visual aid, and should not be interpreted as data-driven network communities. We find that not only are the stereotyped *occupation* concepts in “gender-specific” regions of the semantic network but in many cases, the professions are directly connected to one or more of the gendered concepts, which lends direct evidence of the gender bias given the closer proximity of gender-biased concepts.

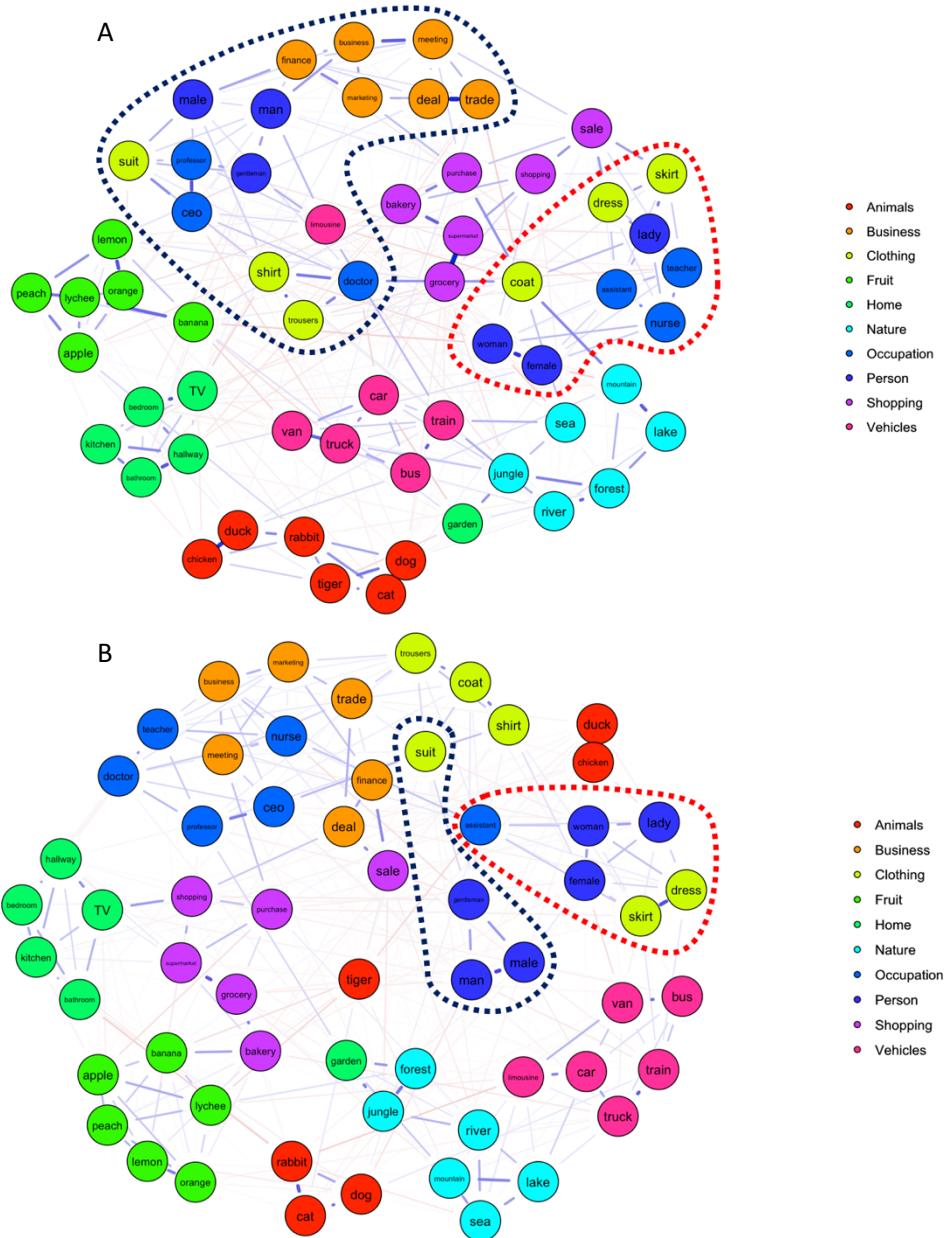


Figure 7.5: Visualisation of the regularised semantic networks by applying a *graphical LASSO* on the network associations grounded in *Google Images* (A) and *Getty Images* (B). The dashed blue (male) and red (female) outlines are manually overlaid to help highlight the male and female regions. Edge thickness depicts the strength of associations between connected concepts.

In the regularised *Getty Images* network, the gender bias is almost absent. The gender-specific concepts of *male*, *man* and *gentleman* as well as *female*, *woman* and *lady* form two distinct “gender triangles”, mostly independent of the occupational concepts. The *occupation* concepts are meaningfully grouped with the *business* concepts like in the unregularised version. However, the concept *assistant* is still closer to the female concept

triangle and is strongly (thicker edge) connected to both the terms *female* and *woman*. This is equivalent to the topology of the unregularised network. Interestingly, the *clothing* concepts are asymmetrically grouped by gender, where the items *skirt* and *dress* are connected to *female*, *woman* and *lady*, and part of the same network region. However, that is not the case for items *shirt* and *trousers*, which are in a more neutral area, although the word *suit* is still directly linked to the word *gentleman*. Therefore, our *scene2vec* trained on *Getty Images* is visibly less biased for gender-occupation, based on our network visualisations. A cursory glance at a sample set of images for the concept *nurse* (see figure 7.6) also reveals a stark contrast, in which the *Getty Image* photographs are typically more professional and gender-neutral (e.g. including male nurses), while those from *Google Images* are somewhat more focused on attractive female models and stereotyped depictions of women smiling.



Figure 7.6: A depiction of a small and quasi-representative¹⁶ set of images for the search term *nurse* in both *Google Images* (A) and *Getty Images* (B).

In our results, the unregularised and regularised networks are similar. For example, like in the unregularised *Getty Images* *scene2vec* representation, in the regularised version we also find a continuous path across the various occupations transitioning from *women/female* to *assistant*

¹⁶ Some images were excluded from this collage, but not *scene2vec*, as these depicted an explicitly adult theme.

leading to *gentleman* via *suit* and *CEO*. Moreover, the intermediary connecting concepts are on a stereotypical gender-biased continuum, with roles like *doctor* and *professor* being closer to the male end, while professions *nurse* and *teacher* are closer to the female. These findings tentatively support that even simple thresholding of the *correlation matrix* to generate the *adjacency matrix* provides replicable semantic network structures.

7.3.3.5 Quantifying Gender Bias

In our investigation of gender bias of *scene2vec*, we go beyond qualitatively analysing network topologies and quantify the bias more rigorously using network analysis. We computationally operationalise this by exploiting the immediacy of concepts as a function of the number of network edges that are traversed using the shortest path between concepts. The prediction is that there will be concepts closer to both genders (e.g. *clothes*), but they do not constitute overt gender stereotyping, while stereotyping occupations is an example of undesirable gender bias. Based on the regularised *Google Images* network, because it is more explicitly gender-biased, we select 19 concepts, including six *occupation*, *business* and *clothing* concepts as well as the word *limousine*, due to its strong male bias. We define *occupation* and *business* concept categories (plus *limousine*) as inappropriate for gender differences, whereas *clothing* is considered an appropriate form of *gender differentiation*, as opposed to bias per se.

To perform our analysis, we use an *unweighted breadth-first search algorithm* (West, 1996) to compute the distances between the 19 concepts and the six gendered concepts (*male*, *female*, *man*, *women*, *gentleman* and *lady*). The average distance is computed across three gender concept variants, resulting in a single composite distance metric for each gender-concept pair. These distances are then scaled across all concept-gender pairings, such that if a concept, is equidistant from, on average, the *male* and *female* concepts, then the bias index will be 1 (parity), indicating no gender bias. However, if the proportion of *female distance* is greater than the

male distance for a given concept, then this indicates the concepts are favourably biased towards *men* (men are closer to the concept). On the other hand, if the *male distance* is greater than the *female distance*, then the concepts are favourably biased towards *women*, as revealed by the closer proximity of these concepts to the trio of female-gendered concepts.

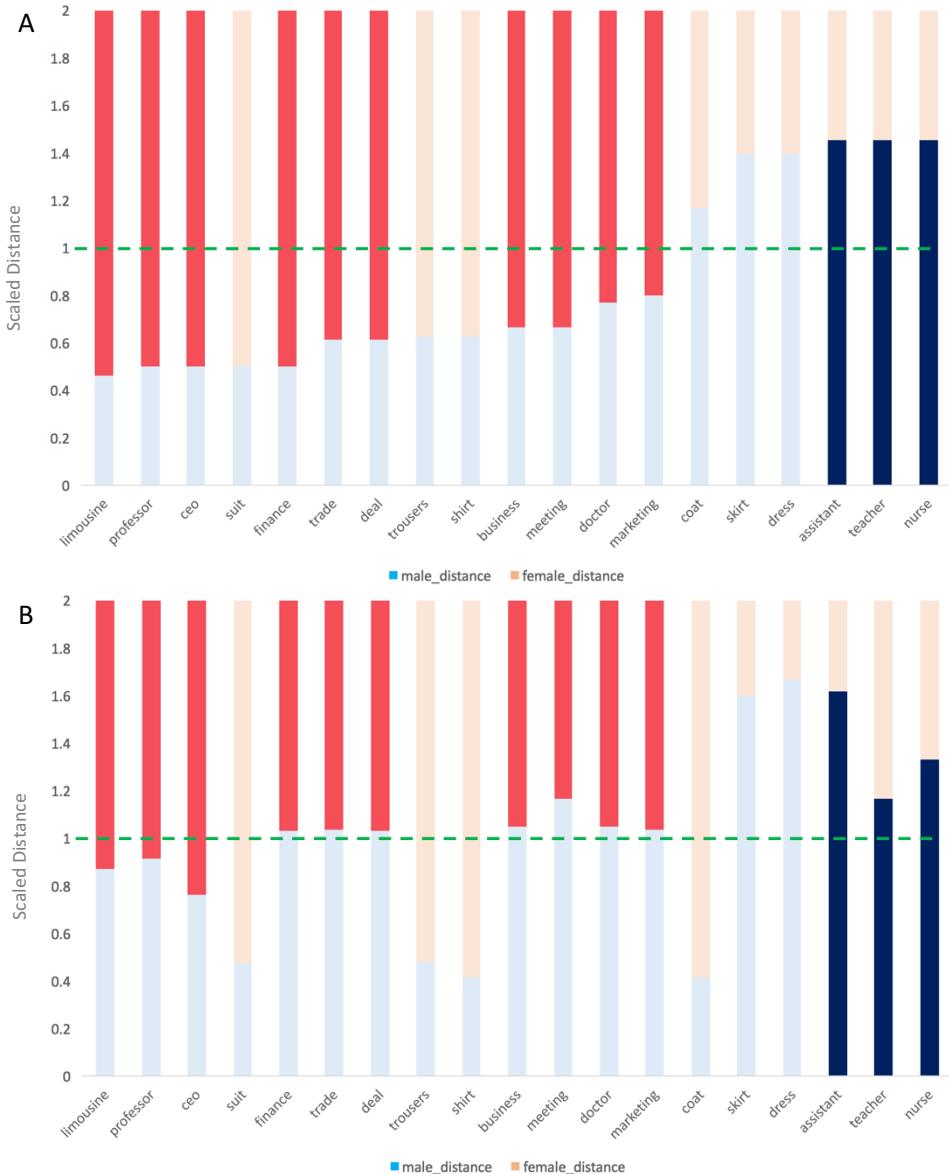


Figure 7.7: Quantifying the gender bias in the semantic networks grounded in *Google Images* (A) and *Getty Images* (B). In both plots, the y-axis *gender bias index* is scaled to compare these two networks with different topologies. Gender-neutral concepts are indexed at 1 (green dashed line), and larger female distances depict bias towards masculine concepts, while larger male distances towards feminine concepts. Darker shades of the colours are used to depict stereotypes/biases of interest.

The results (see figure 7.7) of our networks' quantification reveal strong gender biases for all concepts of interest in the *Google Images*

scene2vec representations. The gender biases, which we deem inappropriate (e.g. *CEO*), are highlighted with brighter colours, while those we feel are acceptable (e.g. *skirt*) are presented in lighter shades. When the proportion of gender bias is non-existent for a particular concept, the *scaled distance* is very close to one, indicating a 1:1 gender ratio (equidistant). Surprisingly, when *scene2vec* is grounded in *Google Images*, the male-leaning bias of *limousine*, *professor*, *CEO*, *finance*, *trade* and *deal* is even stronger than the bias for *trousers* and *shirt*. Similarly, the female-leaning bias for *assistant*, *teacher* and *nurse* is stronger than the bias for *skirt* and *dress*. For the occupations *nurse*, *assistant* and *teacher*, the level of gender bias is equally strong as demonstrated by the equidistant ratios for these concepts. When quantifying the bias in the network using our *scaled distance* metric, the *Google Images* network is biased across all *business* and *occupation* concepts, as well as the word *limousine*. However, the same analysis performed on the *Getty Images* *scene2vec* network reveals fewer *inappropriate* but weaker gender biases, such as male-leaning *CEO* and female-leaning *assistant*, *nurse* and *teacher*. Appropriate gender differentiation of *clothing* concepts is still present and comparable to that of the *Google Images* network. This difference in the *Getty Images* *scene2vec* associations simultaneously demonstrates gender differences for more appropriate concepts, while not for the majority of inappropriate gender concepts, which suggests that *scene2vec* can successfully encode sufficiently nuanced associations.

7.3.3.6 Debiasing

Here, we investigate some “late-debiasing” of semantic associations. In order to run a proof-of-concept version of debiasing, we develop an algorithm called *semantic feature neutralisation* (SFN). We apply this method to the *scene2vec* representation trained on *Google Images* since this is the more gender-biased network. In our SFN method, we start by creating a “feature filter” through which the semantic network is generated. We start with the hidden unit representations of the *scene2vec* neural

network. However, we do not generate correlations directly from these hidden layer representations, but instead, we run factor analysis on the hidden neuron activations and automatically extract all the factors with an *eigenvalue* ≥ 1 .

From the eight factors extracted, we identify the two factors with the highest loadings on the averaged *male* (man, male, gentleman) and *female* (woman, female, lady) concepts. Our SFN method rests on the assumption that these two factors are *allowed to* meaningfully vary for appropriately gendered concepts such as *clothes* but not for *occupation* or *business* concepts¹⁷. We computationally realise this by replacing the *i*th factor and the *j*th concept *factor score* of the *8 factor* \times *60 concept* matrix for the two gender-dominant factors, with a row-wise average of the factor scores across all neutral concepts (e.g. *banana*). Then the following *factor* \times *concept* matrix is transformed into a 60×60 concept correlation matrix, on which we subsequently run the graphical LASSO regularisation and network visualisation, with the same parameters as outlined earlier in this chapter.

When analysing the bias proportion, we use the same set of 19 concepts as used in our previous comparisons, in order to allow direct comparability of the debiasing results. These 19 concepts contain two sets of six inappropriately biased concepts (*occupation* and *business*) along with six appropriately gendered concepts (*clothes*). The presence of two sets of inappropriate concepts is essential, because we predict that our simple SFN method should be successful in *selectively* reducing the level of bias at a network level.

In this experiment, we have three distinct conditions, with the control being the original *Google Images* network shown in *figure 7.5A*, along with a visualisation of the proportion of gender bias plotted in *figure 7.7A*. The three separate conditions allow us to evaluate *bias reduction* for the

¹⁷ This is a challenging and highly subjective notion, which we discuss in section 7.4.

desired set of concepts and *bias maintenance* for either inappropriate (but not treated) or relevant concepts. Our three experimental conditions are as follows: (i) only debiasing *occupation* concepts *{professor, CEO, doctor, teacher, assistant, nurse}*, (ii) only debiasing *business* concepts *{finance, deal, trade, business, meeting, marketing}*, and (iii) debiasing both *occupation* and *business* concepts. We evaluate the network structure and gender bias in an identical way to our previous analysis of *Google Images* and *Getty Images*. We predict that the bias will be selectively reduced, such that in condition one, occupations will be less biased, while all other concepts remain equally biased. Therefore, *clothes*, our untreated “control semantic category”, should have constant levels of bias across all three experimental conditions.

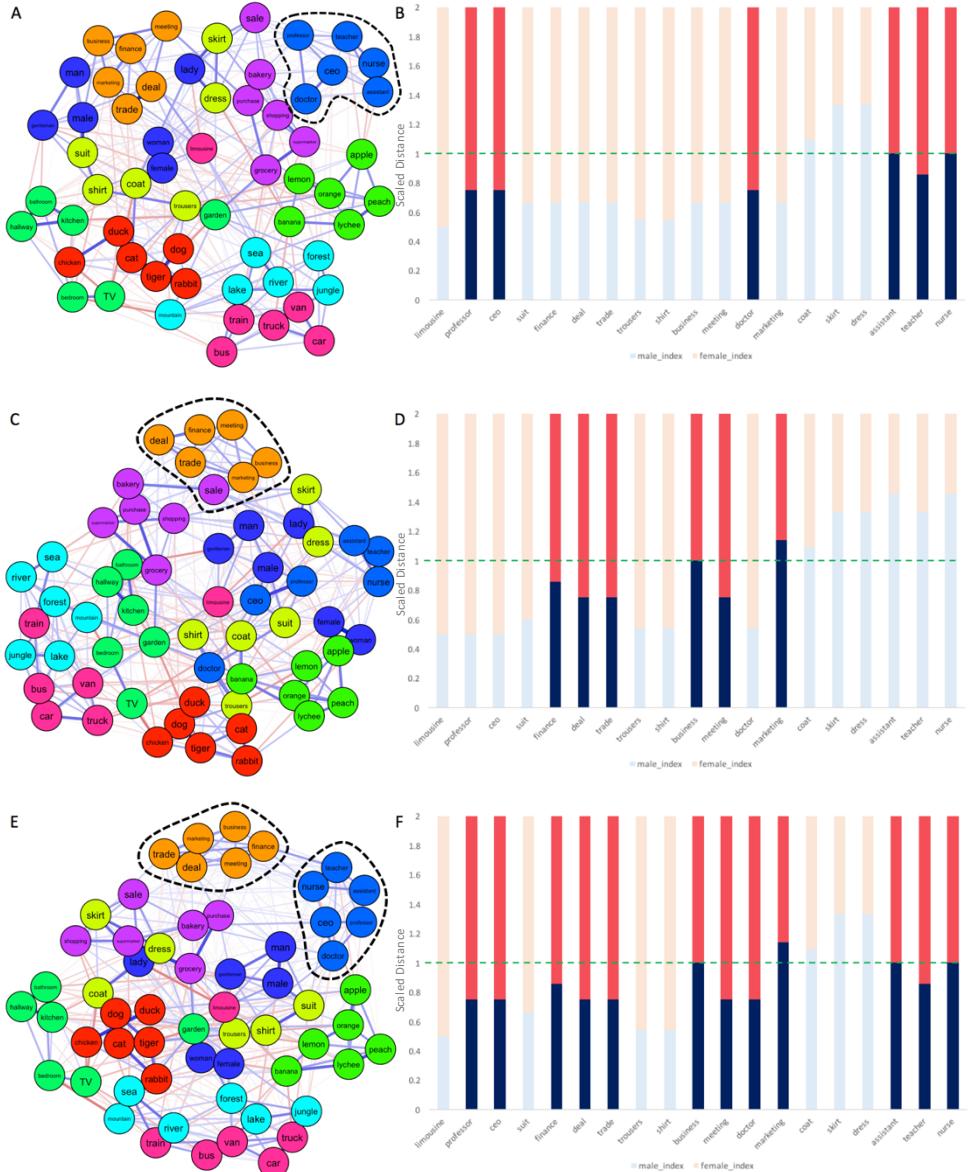


Figure 7.8: The regularised graphical LASSO networks are shown for the conditions consisting of debiasing (A) only *occupation*, (C) only *business* and (E) *occupation* and *business*. The networks have dashed line overlays indicating the specific concepts of interest. The corresponding bias plots B, D and F respectively correspond to the conditions (i) only *occupation*, (ii) only *business* and (iii) *occupation* and *business*. The y-axis represents the scaled *gender bias index*. Gender-neutral concepts are indexed at 1 (green dashed line), and more considerable female distances depict bias towards masculine concepts, while larger male distances towards feminine concepts. Darker shades of the colours are used to depict biases of interest for a given experimental condition. See Appendix D for enlarged images (p. 346 - 348).

In our first condition, applying *selective feature neutralisation* (SFN) to only *occupation* concepts lead to a qualitatively different semantic network (see *figure 7.8A*) grounded in *Google Images*. In this network, there are no longer two distinct network regions split by gender and occupations

as was the case in the original network (*figure 7.6A*). In fact, in this *gender debiased* occupation network, all the occupations are tightly clustered in the same network region, with high levels of interconnectivity, although within this cluster there are two smaller components comprised of the roles *professor*, *CEO* and *doctor* in one, and *teacher*, *assistant* and *nurse* in the other. The topology of untreated concepts is still very similar, such as the grouping of concepts by their categories, and exceptions such as *garden* still preserved in the debiased network, although there are subtle shifts in the visualisation of the topology. The *bias index* proportions of this debiased network in *figure 7.8B* further support our prediction of SFN's *selective debiasing*. Only the gender bias for *occupations* has been re-adjusted, though not entirely, while the biases for *clothes* and *business* concept remain similar to the original network model. However, from *figure 7.8B*, we can also see that not all occupations have been debiased to the same extent. In particular, the more traditionally female-biased roles such as *assistant*, *nurse* and *teacher* (to a lesser extent) are almost gender-neutral, while *professor* and *CEO* have visibly reduced in bias but *doctor* remains equally biased.

In the second condition, where we debias only *business* concepts (see *figures 7.9C* and *7.9D*), the *business* concepts are more tightly bound together in the network with weaker links to male concepts. The quantified bias proportions also show a substantial debiasing effect, with more considerable differences for the concepts *marketing*, *business* and *finance*. In fact, for *marketing*, following the SFN treatment, there is even a small bias toward females. However, for the concepts *deal*, *trade* and *meeting* the gender bias remains preserved, indicating that the two gender factors selectively neutralised via SFN, are unlikely to have played a contributing factor for the bias inherent in *deal*, *trade* and *meeting*.

Lastly, in the third SFN-based debiasing experiment (see *figures 7.9E* and *7.9F*) where we jointly treat the *occupation* and *business* concepts, the network no longer explicitly reveals the strong gender bias of the original, *Google Images* based *scene2vec* network model (*figure 7.5A*). This SFN-treated

network's bias index also reveals that the semantic bias from the first *occupation only* and second *business only* conditions are integrated additively.

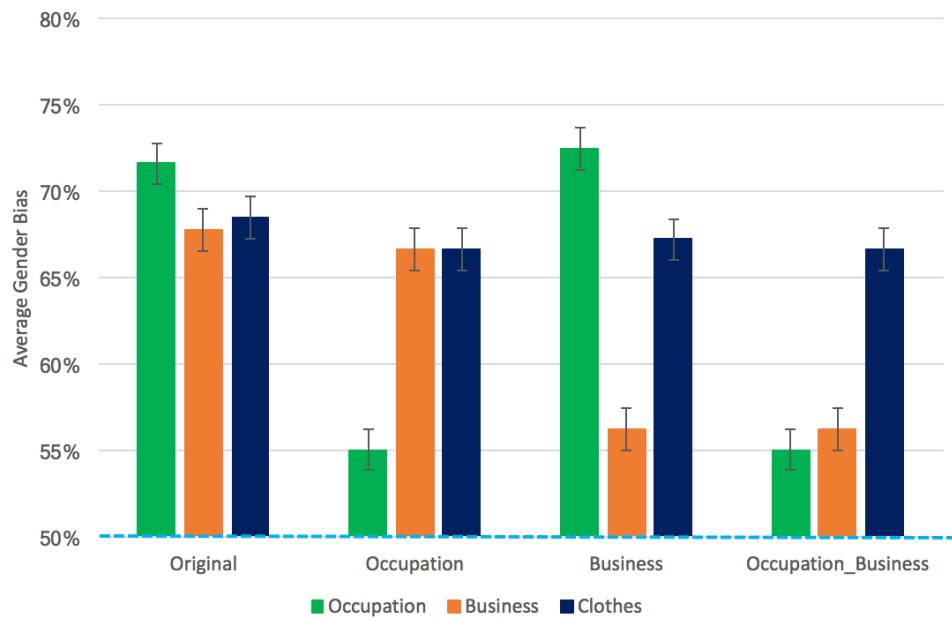


Figure 7.9: A summary plot of average bias levels for the original *scene2vec Google Images* network along with the three debiasing experimental conditions of (i) *occupation-only*, (ii) *business-only* and (iii) *occupation and business*. The averages biases are shown split by the occupation and business categories in addition to the “control category” *clothes*. The average gender bias (y-axis) is calculated within-concept categories, by taking the largest scaled network distance from the averaged *male* or *female* concepts and dividing it by the overall scaled differences across *both* genders. The gender bias quantifies an overall bias proportion across both genders. The blue dashed line (50%) indicates no gender difference. The error bars reflect 95% confidence intervals based on 100 iterations of differently seeded initialisations of the *scene2vec* neural network training.

We summarise our debiasing experiments in *figure 7.9* and outline four key findings of our SFN debiasing treatment. First, the automatic SFN debiasing method can successfully, but not perfectly, debias the resultant semantic network as demonstrated across the three conditions. Second, across all conditions, the bias is never entirely removed but is reduced. Arguably, our results in *figure 7.9*, mask the findings from the *business-only* condition, in which only half of the concepts' biases are reduced. Third, SFN can be used to selectively target and debias individual groups of concepts as seen by the strong interaction between the *occupation-only* and *business-only* conditions. Fourth, appropriately gendered concepts such as

clothes (man:shirt :: woman:skirt) can maintain their preferential gender associations despite targeted (*occupation or business debiasing*) or widespread application (*occupation and business debiasing*) of the SFN debiasing treatment.

7.4 Discussion

In a series of experiments, we demonstrate the capacity for our scene-based representation, *scene2vec*, to capture human-like gender biases successfully. We also embark on an explorative foray into developing a proof-of-concept “late debiasing” technique, called *selective feature neutralisation* (SFN). From scene-based distributed representations capturing gender biases, we argue that there might be a range of empirical and theoretical implications for *computational cognitive modelling*, *artificial intelligence* and perhaps even somewhat tentatively for *gender studies*. In this section, we will start by contextualising our main findings within the broader literature of both implicit cognitive biases from social psychology and *machine learning fairness*. This review leads to the evaluation of the relevance of this work to cognitive modelling of semantics in particular but also AI systems in the real world. We will try to pre-empt some strong criticisms and limitations of our operationalisation of *gender* in this study, particularly regarding *gender fluidity*, which is something the present research unfortunately wholly ignores given our binary gender construct. We examine germane directions for future research, some of which we outline as next steps, while others we propose as open challenges that we currently have no tractable way of solving. Lastly, we discuss the implications of our promising but sub-optimal debiasing results for scalable applications in the real world. We conclude with a broader discussion of gender bias being a cultural construct, grounded in our interdependent sensorimotor and language-based or *symbol interdependent* regularities as opposed to biology.

7.4.1 Contributions

With scene2vec, we conceptually replicate a well-known human bias established in social psychology using both self-reported measures and IAT. Even though the term “stereotype” was introduced in the social sciences almost a century ago, by Lippmann (1922), there has not been a great deal of attention on the subject of bias (Ashmore & Del Boca, 1981). Only half a century after the term “stereotype” was coined, did psychologists start investigating gender stereotypes. One of the most widely cited early examples of empirical support for explicit gender stereotypes originates from Broverman et al.’s (1972) discovery of the more positive evaluation of masculine traits like *competence* than feminine qualities such as *expressiveness* and *warmth*. This finding revealed not only a difference in the associations that stereotype men and women but also a difference in the *appraisal* of these gender biased attributes, irrespective of sex, age, marital status and education.

More recently, Rudman, Greenwald and McGhee (1996) reliably documented *implicit* gender biases associating men with *power*, as represented by concepts such as *CEO*, *doctor* and *professor* in our study, and women with *nurturance*, which some might argue is traditionally seen as less related with power, and more with professions like *nurse*, *assistant* or *teacher*. Furthermore, Rudman and Phelan (2010) showed the presence of strong gender roles (e.g. a *male surgeon* and a *female nurse*) and also that when female participants were primed with such stereotypical gender roles, this led to an increase in gender-biased stereotypes and career aspirations, relative to controls, and also mediated their significantly lower interest in traditional masculine occupations. Even more surprisingly, when female participants were primed with non-traditional gender roles such as a *female surgeon* and a *male nurse*, this lowered their self-concept of leadership in conjunction with also lowering their interest in traditionally masculine professions. Therefore, Rudman and Phelan (2010) discuss that the priming of both traditional and non-traditional gender roles leads to

greater gender bias, through two separate mechanisms, in the first case through activation of stereotypes while in the second scenario, through the reduction of self-concepts. We believe that this research is relevant to contemporary discussions of gender biases and how to overcome them, given that this evidence suggests the task to be particularly challenging. It also highlights that potentially well-meaning and supportive initiatives like highlighting non-traditional roles can do more harm than good concerning achieving professional gender equality in society, which is increasingly shaped by a plethora of machine learning algorithms.

Given the rise and ubiquity of big data in the contemporary commercial, but also, academic sector, it is only understandable that public scrutiny over data analytics has been steadily increasing (Bollier & Firestone, 2010; Richards & King, 2014). In the recent past, there has also been more attention on the generation and legislation of *big data guidelines* (e.g. Zook et al., 2017) as well as the deployment of *privacy-preserving* algorithms (Hunt et al., 2018). However, the debate on machine learning fairness, in particular, related to gender stereotyping, has lagged behind considerably given the widely perceived lay-perception of machine learning “being fair” because “it is mathematics”. We outlined the computational language-based experiments of Caliskan et al. (2017) using the synthetic equivalent of the IAT, the *Word-Embedding Association Test* (WEAT), which replicates all of the original implicit biases, including gender bias. Our computational results build on this and demonstrate that gender biases are also prevalent in our scene-based semantic representations.

Our grounded *scene2vec* representation introduces biases in the modelling of semantic cognition. One could argue that our introduction of more ecologically valid semantic representations is counter-productive and perhaps even harmful because traditional feature-based approaches (e.g. Rogers and McClelland, 2004) might not have the same problems. We strongly disagree with this argument. First, feature-based connectionist

models, to the best of our knowledge, have not been constructed in cognitive science (or another field) to evaluate phenomena such as gender bias. Second, although theoretically our grounded findings could be replicated using modeller-defined feature sets, this introduces the dilemma of the computational model merely generating “associations” from desired patterns of “cognitive phenomenon” that the modeller determined *a priori*. Third, it is difficult to appreciate the theoretical advantages of exploring biases in networks determined by non-data driven inputs.

7.4.2 Importance of Grounded Data

We propose that one of the aims of cognitive science should be the acquisition and storage of raw perceptual and linguistic information simulating human exposure to relevant stimuli. Ideally, this could be achieved in an ecologically valid and developmentally sensitive manner. This would allow for the simulation of concept acquisition from a constructivist semantic network perspective during the early years of human development. Simulating the later years of human development, one might even investigate pathways of semantic memory degradation, in typically ageing cohorts but also those afflicted with conditions such as *semantic dementia*. However, in our view, much of cognitive modelling has predominantly focused on specification details of the underlying mathematical, statistical or computational assumptions (e.g. *model-based* or *model-free reinforcement learning*, *MDP* or *POMDP*, *connectionist networks* versus *singular value decomposition*). This comes at the expense of using simple hand-crafted toy datasets. This “data” can consist of either hand-crafted binary features based on verbalised feature norms or even generated entirely randomly with certain key requirements such as *proportion overlap* required between concepts within and between a range of categories (e.g. Hoffman et al., 2018). This is an “inverse case” instance of the infamous computer science acronym *GIGO* (*garbage in, garbage out*), which states that if flawed data enters into a computational process, flawed

outcomes will naturally follow. We suggest that for many cognitive computational models, since a great deal of effort is spent crafting the input representations to demonstrate a particular phenomenon, it is unsurprising that the outputs then are meaningful based on *a priori* meaning-laden assumptions of the modellers.

At this point, we would like to explicitly foreshadow one of the central themes of not only this chapter's findings but those from some of our modelling experiments from chapters 4 and 5 as well. In the words of Cooper and Shallice (2006, p. 892), we label the theme "the importance of the training set". We selectively illustrate an interesting exchange, spanning years, between two positions espoused initially in Cooper and Shallice (2000) and Botvinick and Plaut (2002). We will avoid going into the details of the modelling, as the high-level summaries will suffice for our purposes of highlighting *why* training data is more critical than the algorithmic specifications per se. Clearly, in doing so, we are also influenced by the similar arguments conveyed by Louwerse (2011) in the context of language-based statistical regularities.

Cooper and Shallice (2000) developed a novel computational model for contention scheduling and control of routine actions, albeit, with a hand-crafted set of schemas. Botvinick and Plaut (2002) critiqued this hand-coding of schemas and instead provided a connectionist alternative, which could replicate many of the same action selection patterns and more importantly, capture a range of common *action slips*. However, as outlined in Cooper and Shallice (2006, p. 894), Botvinick and Plaut did not solve the problem of hand-coding the hierarchical schemas, but simply hand-coded the relevant training data in order to generate the required learned associations. Moreover, although Cooper and Shallice (2006) did concede that if these representations could be determined through the observation of human action sequences, or in our terminology, were "grounded", then that might be interesting. Therefore, in both Cooper and Shallice's (2000) original symbolic model and Botvinick and Plaut's (2002) version, the

differences in *model type* are less critical than *what* is being hand-coded - the hierarchical rules or the training data used to feed a recurrent neural network to learn these hierarchies. In our research, however, we use freely-available image data as the source of all our computational semantic experiments. Thus, underlying semantic nuances encoded in the hidden neurons of our various *Perceptual Scene Vector* (PSV) or *scene2vec* models are of practical interest as these are not specified *a priori*.

Our position claims that *grounded representations* are based on real-world stimuli and can reflect genuine and non-trivial differences when grounded in different scenarios. In this study, we tested our prediction of *Google Images* being more gender biased concerning stereotyped gender occupations and business concepts than *Getty Images*. There are differences between *Getty Images'* repository, based on a strict code of conduct and editorial guidelines, versus, *Google Images'* minimal filters. Using grounded representations trained on a diverse set of real scenarios, like in this study, we can generate and evaluate semantic biases that are not determined by the modeller, but informed by a variety of different environments.

Therefore, our grounded semantic models can be more readily subjected to actual analysis in order to compare with semantic topologies extracted from human participants. This potentially unlocks opportunities for cognitive computational models to connect with a range of psychologically more interesting problems as opposed to highly simplified caricatured meaning models of prototypical *birds, fish, mammals, trees* and *plants*. Simultaneously, we are cautious not to dismiss the decades of critical and highly influential connectionist models of semantic cognition based on toy models. Much of our present thesis is solidly grounded and inspired by this research. We propose a synthesis of both *theoretical toy models* (e.g. Rogers & McClelland, 2004; Hoffman et al., 2018) in conjunction with larger-scale semantic models based on the automatic or semi-automatic extraction of meaning from the real world. The first set of toy models could

inform a more *information theoretical* perspective of cognitive science and define the core limits of processing mechanisms.

On the other hand, the more *applied cognitive models*, experimenting with different real-world inputs, could deploy scaled versions of the toy models in a diverse range of psychologically interesting domains, like in our case with gender bias. If cognitive modellers pursue both the theoretical and applied cognitive models, then through mutual reinforcements and constraints, we will gradually further our mechanistic understanding of cognitive semantics. We currently lack large-scale cognitive semantic models that can simulate upwards of tens of thousands of concepts, reaching human-level conceptual system complexities.

Cognitive science has been critical to the success and development of artificial intelligence, given the co-evolution of various paradigms in both disciplines. However, more recently, the present author tentatively suggests, there has been a widening chasm between the two disciplines. In the fields of *machine learning* (ML) and *artificial intelligence* (AI), there has been a steady progression towards using more data, more computing power and increasingly sophisticated statistical learning algorithms. We acknowledge the highly distinct objectives of both fields. Cognitive modelling is focused on mechanistic, algorithmic and implementation-level aspects of human cognition. On the other hand, AI/ML adopts an engineering mind-set for building practical, intelligent systems, irrespective of their biological and cognitive micro-foundations. Nonetheless, in our view, cognitive modelling has all-too-often focused on overly simplistic models with highly fine-tuned sets of model parameters, as opposed to relying on gathering high quality and cognitively plausible data. This is one of the gaps in the extant literature on semantic cognition that we are aiming to bridge with our current work while investigating semantic gender biases.

7.4.3 Gender Bias

Gender bias is a significant societal problem, that is not simply undesirable and politically incorrect but also has severe economic and welfare benefits. Economic analysis has shown that *gendered evaluations* of occupational characteristics (e.g. women as nurturing) have a real impact on the *gender pay gap* (Kilbourne et al., 1994). In real terms, given that female full-time employees in the US earn approximately 79% of what their male counterparts earn (Blau & Kahn, 2017), we extrapolate these findings, on the basis of average current annual salaries, to suggest a gender pay gap of circa \$10k US dollars for full-time employed women. However, these statistics mask the real extent of the gender pay gap given the comparison of like-for-like women and men in full-time employment. When taking into account the disturbing reality of increased psychological and physical vulnerability of women who are not financially independent, see Sanders (2015) for a qualitative investigation, the severity of the gender pay gap becomes all too apparent. Therefore, understanding and helping to overcome gender biases is of high importance.

A more detailed understanding of our environments coupled with fostering a culture without overt biases can start to accelerate the battle against this costly inequality impacting approximately half the world's population. A straightforward, actionable insight of our present work would be to recommend the use of editorially-curated content like the photographs from *Getty Images*, given the biases of search engines like *Google Images*. Future research questions might explore differences across other search engines or image repositories from providers with a different editorial leaning to *Getty Images*.

7.4.4 Limitations

Despite our best efforts of investigating a grounded perspective on scene-based semantics, our 60 concepts chosen were modeller-defined.

Therefore, a reasonable criticism of our current investigation is the inclusion of the six gendered terms used (*male, female, man, woman, gentleman, lady*). Although a valid critique, we avoid the use of concepts such as *he/she* because many of the search results returned are that of animals/pets. However, future research ought to investigate a broader range of concepts than the 60 used in this study. Our limitation was mainly shaped by the objective of having network visualisations that were relatively small in network diameter to increase interpretability.

A more severe limitation of the present research is the use of polarising gendered concepts in order to investigate gender biases. We feel this is a necessity for operationalising our research hypothesis. However, more recently, both in popular culture and for a more extended period in the academic studies of gender (e.g. Ortner & Whitehead, 1981; Chodorow, 1995; Parker, 2016) it is no longer acceptable to mention two distinct genders, as was done in the present research. There is even some disagreement on the existence of two *biological sexes*. However, this is beyond the scope of the present discussion, see Fausto-Sterling (2012) for a detailed review.

Nevertheless, there is a consensus that “[g]ender reflects [a] sense of self, social expectations, and role behaviours”, (Parker, 2016, p.165), which have evolved from the dichotomy between *men* and *women*. This notion of *gender fluidity* has gained a great deal of support and attention in recent years and is very likely to continue doing so. Intriguingly, this notion of the dynamicity of concepts is the focus of chapter 6 of this thesis. Therefore, the notion of concepts being dynamically constructed as a function of cultural pressures is very much in line with our work in the present thesis. However, unfortunately, at this stage, the present author is unaware of conducting investigations into gender biases, while accommodating for gender fluidity, due to the need of operationalising gender in order to model relevant statistical regularities from the environment. Perhaps, in the future, there might be a range of *gender*

identifiers for the continuum between *male* and *female*, although even that will have classification challenges given that individuals might differ based on their position on this spectrum as well as the stability of their culturally constructed gender identity. Furthermore, identifiers also go against a fluid gender continuum.

The source of images used for meaning representation is critical, as revealed by some of the stark contrast in gender bias between models trained on either *Google Images* or *Getty Images* collections. There is a high risk of reproducing biases present in either historical data or the present cultural discourse. Therefore, in particular from an AI/ML perspective, as opposed to cognitive modelling per se, it is beneficial to understand if the semantic networks can be debiased successfully in a targeted and systematic fashion. Debiasing algorithms and machine learning fairness is a new and thriving discipline (e.g. Joseph et al., 2016). However, many of these cases are based on adjusting for class imbalance using statistical and mathematical techniques. In our case, gender biases in semantics, the biggest challenge we face is that tried-and-tested rebalancing of input data not only reduces bias but also distorts the semantic space itself. Furthermore, the question we need to ask ourselves is although gender bias is undesirable, are we distorting the space so much with debiasing that it loses resemblances to human meaning topologies? If yes, the resultant semantic network loses its original purpose.

7.4.5 Debiasing Semantic Networks

The current debiasing experiments support the feasibility of simple dimension-based filtering techniques like our *selective feature neutralisation* (SFN) to target and reduce gender bias in our grounded *scene2vec* representation. As intended, SFN preserves much of the remaining network topology. However, SFN also successfully preserves gender-appropriate references. This is particularly important for cognitive and ML/AI models of semantics, because we argue that there is no clear

distinction between *bias* and meaning-related cues, as these two phenomena are synonymous - *experience-based statistical regularities*.

We propose a tripartite framework of characterising types of techniques for debiasing representations, consisting of (i) early (*data*), (ii) middle (*algorithm*) and (iii) late (*adjustment*). In the first class, the underlying data is selected from a particular source or is adjusted (e.g. weighted) in order to either reduce or eliminate biases before any modelling. An example would be choosing a less biased data source such as *Getty Images*, as confirmed by our study because it has a strict professional set of editorial guidelines that most online image repositories lack. Alternatively, others have demonstrated the successful use of data pre-processing for increasing *machine fairness* (Calmon et al., 2017).

The second method of debiasing (middle/algorithm-level) consists of modifying the underlying learning algorithm to reduce some objectively defined bias criteria. Solutions that fall into this category would be, for example, direct modification of the objective loss function, which impacts the loss being backpropagated through the layers of a neural network. Finally, the third class consists of applying an adjustment to an already trained distributed representation. Our *selective feature neutralisation* (SFN) is an example of this last category given that the feature filtering applies an adjustment to the learnt associations to reduce bias.

We now present an argument as to why there is an *ease of implementation* hierarchy of our three classes of debiasing for real-world applications on large-scale AI / ML architectures. In order to focus specifically on semantic associations, we will select a particular class of widely applicable algorithms called *recommendation engines*. At their core, recommendation engines, irrespective of their symbolic, sub-symbolic or hybrid architecture, ingest large volumes of continuous streams of data from users (e.g. web-based clickstream records, cookies or social media likes) to generate higher-dimensional associations, which are then clustered by ever-changing business strategy and other contextual factors (e.g.

weather). This high-dimensional space can represent customers' behaviours and (to an extent) the latent needs, which is similar to distributed semantic models. It is also worth noting that real-world recommendation engines, like Sky's¹⁸ movie recommendations, are based on more than a dozen separate machine learning solutions, each ingesting real-time data from websites, app usage information, social media profiles, customer research and most critically, the actual viewing habits of the customers. Typically, the deep learning algorithms deployed for constraint-based optimisation are operated on ML stacks ranging from two or three models to dozens of distinct models, each with specific loss functions. Prototyping, productionising and maintaining these models requires a variety of programming languages/architectures such as Python, R, C++ or even SAS, GCP, Hadoop and Apache Spark. Lastly, the final layer of complexity stems from various machine learning solutions continuously operating across 24+ million customers' data records across national borders, spanning a wide range of different data processing legislation and customer-specified preferences.

We argue that late debiasing solutions are particularly desirable for the debiasing treatment of semantic networks and real-time recommendation engines. Our SFN method is a "late technique", that avoids time-consuming and costly efforts of having to collate a well-balanced dataset / iteratively weight original data or make algorithm-level adjustments which require a great deal of dedicated *research and development* (R&D) time on a problem-by-problem basis. To the best of our knowledge, it is a highly intractable problem to assume that middle/algorithmic level debiasing can be successfully executed given the complex ecosystems these AI/ML algorithms operate within commercial environments.

¹⁸ Sky is Europe's largest pay-tv and media business spanning the UK, Germany, Italy and Austria.

Selective feature neutralisation has the advantage of not requiring data augmentation or re-weighting, which in many real-time systems would be difficult to realise. Furthermore, *neutralising* core features at the *end-stage* of an algorithm ensures that various debiasing (e.g. *gender*, *ethnicity*, *age*, *affluence* or *region*) requirements can be dynamically met by identifying critical underlying semantic features with specific debiasing targets as performed in our analysis. One interesting avenue for future research might be to investigate the relative quality, through comparisons with *ground truths*, of *early*, *middle* and *late* debiasing techniques, along with metrics of implementation and computational complexity, as the number of intertwined debiasing target groups (e.g. *gender*, *ethnicity*, *sexual orientation* and *affluence*) increases. We predict that for early- and middle-stage debiasing, quality will decline and complexity will increase, while for late-stage techniques, both quality and complexity are likely to remain constant.

Ensuring that debiasing solutions can scale effectively is critical from a data science and information technology perspective. Efforts to promote debiasing for enterprise-wide systems such as *automated curriculum vitae* (CV) association mining, requires large companies, in media, technology and finance to be able to debias their AI solutions with ease and transparency. With our development of the simple *selective feature neutralisation* (SFN), we aim to be able to offer one variant of this class of solutions that are easily deployable and could dynamically adapt to the changing requirements of debiasing products and services while still preserving appropriate gender differences.

In addition to technological challenges, sociological and cultural factors cannot be side-lined. Meaning is context dependent, culturally attenuated and changes over time. Furthermore, in broader machine learning fairness efforts, we might easily reach a consensus on deploying debiasing algorithms similar to SFN, in instances where a job recommendation engine *should* surface job adverts irrespective of gender.

In other instances, there might be uncertainty and public disagreement. For example, does an association rule-based recommendation system have to recommend products in a gender-neutral way even if behaviours suggest shoppers would not be interested? Dilemmas would arise if some supermarkets adopted a gender debiased solution and lowered their ad targeting utility, while their competitors opted not to debias their algorithms and benefited from higher ad targeting utility, revenues and profits.

Despite our proof-of-concept demonstration of SFN and the advantages of late-debiasing methods, some questions regarding the efficacy of SFN remain. For example, would SFN be equally successful at reducing gender bias in a larger scale model (100s of random concepts) or even a human-level conceptual system (10,000+ concepts)? We predict that SFN is likely to perform poorly in slightly larger-scale “toy systems” but perform better on significantly larger/human-scale synthetic cognitive systems. Although throughout this thesis, we constructively critique hand-coded feature selection, we too have inadvertently “constrained” our set of 60 concepts. We opt for categories of concepts that are predominantly not related to gender biases (e.g. *fruit*, *home*, *nature*, *vehicles* and *animals*) but others that are likely to be more related (e.g. *occupation* and *business*) so that we can investigate gender bias. However, by constraining the stimuli to these 60 concepts, we have ensured that the concepts’ semi-automatically grounded features are more likely to be constrained such that we find interpretable overlaps between various concept categories. In other words, if we pre-determined the six gendered concepts (*man*, *woman*, *male*, *female*, *gentleman*, *lady*) and selected the remaining 54 concepts at random from a lexicon, then it would be unlikely for us to replicate our current findings. It is likely that we would have to sample a large set of random concepts until we have a sufficiently large number of meaningful comparators like the concepts in our *occupation* or *business* categories. Therefore, we predict that SFN is more likely to show promising results in smaller (and somewhat

curated) datasets like the ones used in this study and significantly larger datasets with 10,000+ concepts.

Furthermore, in our implementation, SFN extracts and selectively “neutralises” two dimensions respectively corresponding to *male* and *female* concepts. However, even in our small-scale semantic model, we find concept instances (i.e. 50% of *occupations*) that are not successfully debiased. We suggest¹⁹ that by only “neutralising” a subset of factors associated with gender concepts we ignore other factors that make smaller contributions. Therefore, in future studies, we aim to explore the impact of varying the number of dimensions to which SFN is applied. Perhaps, there are likely to be domain and bias-specific optima. This might also shed light on the extent to which grounded semantic dimensions have a “long-tail” problem of bias. We conjecture that much of the bias might be captured by a small set of main dimensions, while achieving an optimal “bias-free” representation might require SFN to be applied to a more substantial proportion of weaker/long-tail dimensions. However, in this case, SFN would need to be adapted as otherwise there is a risk of equally neutralising too many concepts across numerous dimensions. We conjecture that some form of weighting might need to be introduced to up-weight the debiasing of particular dimensions and concepts, while down-weighting concepts and dimensions only slightly impacted. Perhaps the *weighting factor* might be proportional to the different dimensions and concepts bias exhibited.

Another limitation of our study is the lack of an objective metric outlining the *quality* of semantic representations as a function of debiasing (e.g. original network, debiasing either *occupation/business* or both *occupation* and *business*). Qualitatively, we feel that the original *Google Images* network is more meaningful, especially when compared to the debiased alternatives. We do not advocate the perpetuation of gender

¹⁹ Our exploratory post-hoc analyses of the *factor* × *concept* matrix and *scene2vec* features confirms this.

biases. However, given the well-documented implicit cognitive biases discussed in the introduction, it is not surprising that a biased network might appear to be more meaningful to human observers. A future conceptual replication of this study with a broader range of occupations and business concepts could also be concurrently run with human participants using the cognitive dimensions-based approach from chapter 6. Alternatively, a language-based distributed representation such as LSA or GloVe could be used as a linguistic comparison. This would allow for direct quantitative and qualitative comparisons to ever-increasing levels of SFN debiasing.

7.4.6 Summary

We believe that a world without harmful biases is a goal worth striving for, not only for ethical and moral but also economic factors. In this study, we provide the first ever computational demonstration of human-like stereotypical gender-occupation biases acquired from real-world visual scenes. Furthermore, our scene-based semantic representations are sufficiently sensitive to represent different levels of gender bias depending on the source of photographs used for model training. Like Caliskan and colleagues (2017) demonstrated for language, we show that in grounded visual scenes, our scene2vec can also automatically acquire what humans know implicitly, including undesirable biases. This opens up new avenues for future research on mechanistically investigating the culturally-attenuated sources of gender (and other) biases. Future cognitive modelling research might even track how semantic networks change over time as visual discourses mature. In the process, we have also shown the effectiveness of regularisation techniques applied to cognitive semantic network models. The aim is to increase replicability of semantic networks and to minimise the over-interpretation of spurious associations, especially as networks scale.

Lastly, our findings lend empirical support to qualitative ideas expressed by gender theorists regarding gender bias not being a consequence of biology, but rather a direct consequence of our culture. In our view, cognitive modelling using a grounded perspective will help build bridges between the cognitive sciences and other disciplines investigating human belief systems. We hope that our present computational studies will lead to more applied investigations of essential topics such as gender bias in the field of computational cognitive science.

Chapter 8

General Discussion

8.1 Abstract

This thesis supports the recent trend in artificial intelligence and cognitive science to move towards the processing of raw sensory information, resurrecting ideas from Brooks-style situated robotics. Our original contribution to cognitive science focuses on modelling semantics without recourse to hand-engineered features, by grounding meaning in real-world visual scenes. The computational and behavioural studies of this thesis support the hypothesis of highly dynamic or context-specific meaning topologies grounded in the visual messiness of the real world. In this closing chapter, we start by outlining our main contributions. Then, we address Dreyfus's (2007) core challenges to grounded AI as well as other viewpoints like the symbol grounding problem and discuss prospective research and commercial implications of grounded semantics. We conclude by exploring three future research opportunities, which are: (i) creating a *ground truth* for benchmarking cognitive semantic models, (ii) large-scale semantic modelling, and (iii) developmental investigations of semantic

network topologies. Lastly, we propose that a move towards human-level AI will benefit from realistic *virtual grounding*, enabling rapid iterative progress while also providing a novel, ecologically-valid foundation for cognitive modelling of semantics.

8.2 Introduction

This thesis has demonstrated the feasibility of grounding semantics in the real world. Cognitive semantics is a complex and multifaceted phenomenon, which is at the heart of complex, intelligent behaviours. Therefore, semantics is an ideal testbed for investigating grounded perspectives, given the ubiquitous nature of semantic processing across a range of higher-order cognitive domains. The study of cognitive semantics has come a long way since the early theories dating back to Quillian (1967). However, we argue that despite numerous advances in investigating cognitive semantics, such as *symbolic models* (Nagy, Seth, & Stoddard, 1986), *simple neural networks* (McClelland, Rumelhart, & Hinton, 1988; Rumelhart, Hinton, & Williams, 1986), *self-organising maps* (Kohonen, 1990), *recurrent neural networks* (Hölldobler, Kalinke, & Störr, 1999), and more recently, supposedly, “embodied” hub-and-spoke modular neural networks (Hoffman et al., 2018), all of these developments have one limitation in common - they are based on hand-engineered or simulated datasets, with little or no resemblance to the real world.

These models lack ecological validity, although some of them have been successful in narrowly simulating certain behavioural and neuropsychological phenomena of semantic cognition. Nonetheless, contemporary models of semantic cognition have little in common with human phenomenological experiences of the world. The input features used are either hand-coded based on properties revealed from feature-norm studies (Cree & McRae, 2003; McRae et al., 2005) or are merely randomly generated based on set thresholds of overlaps for concepts

assumed to be in the same category (e.g. Hoffman et al., 2018). This thesis, on the other hand, uses real-world visual stimuli to develop an alternative scene-based approach to representing meaning.

The origins of investigating meaning have well-known Western philosophical, linguistic and empirical roots but also overlooked Indian philosophical precursors dating back millennia. The ancient South Asian study of *Samkhya-Yoga* synthesises perception, inference and memory. Our focus has been on the interplay between perception and semantic memory, which is a natural consequence of our grounded approach across both our computational and empirical studies. The extant literature on grounded cognition (e.g. Barsalou, 2008, 2010; Glenberg et al., 2008) suggests an increasingly influential role in emphasising the importance of real-world sensorimotor associations which constitutes the building blocks of human intelligence.

Grounded cognition rejects the notion of cognition being the result of computational manipulation of amodal symbols, devoid of perceptual and environmental content. Throughout this thesis, we have steered clear of grounded perspectives on semantics which overtly oppose *computations* and *representations* (e.g. Chemero, 2011), because we feel this literature typically misconstrues the core tenets of computational cognition and conflates a number of issues due to misinterpretations of mechanistic accounts, which are beyond the scope of our focus on grounded semantics.

In our literature review (chapter 2), we discussed how cognitive psychologists overlook the historical influences of artificial intelligence (AI) and cognitive robotics while only sketching out the importance of these fields for the future of grounded cognition research. We resurfaced some of the pioneering but somewhat forgotten AI and Brooks-style robotics research from the 1970s and 1980s, which have shaped modern-day grounded cognition. In this thesis, we embark on operationalising Brooks' (1999) original vision of *from pixels to predicates*, although we focus more narrowly on the grounded interdependency of associations for meaning

extraction and representation. This has been possible due to recent advances in machine learning and the availability of web-based image repositories, which we argue, cognitive scientists are not sufficiently exploiting, unlike our AI counterparts. Although we use a range of distributed language-based models and incorporate language tags as a grounded input in chapter 7, our focus throughout the thesis has remained on naturalistic scene-based object co-occurrences containing sufficient variability for capturing rich concept-to-concept associations.

8.3 Main Contributions

Our approach to grounding semantics in the real world use a range of deep learning models to process visual inputs and shallower neural networks for representing the scene-based statistical regularities, predominantly focusing on object co-occurrences, although emotional expressions and linguistic tags are later incorporated. A range of dimensionality reduction techniques and visualisations such as multidimensional scaling (MDS), dendograms, correlation plots and regularised and unregularised network analyses are used to explore the hidden neural network representations capturing the statistical associations and emergent generalisations. Our focus has been on this bottom-up concept-level understanding.

Throughout our computational modelling studies, we incrementally developed our approach to grounding. We started with shallow feedforward neural networks processing rudimentary one-dimensional silhouettes of images, which are compared with traditional feature-based and hybrid inputs (both feature-based and grounded) to conceptually replicate the mutually reinforcing nature of hybrid representations (chapter 3). Building on Goldstone and Rogosky (2002), we show that hybrid representations have a markedly slower rate of decline in concept classification accuracy as a function of increasing levels of noise

perturbations applied to the hidden layer representations. Grounded representations perform the poorest, while feature-based inputs are moderately tolerant to increasing levels of noise. This supports a more pluralistic cognitive semantic perspective.

We then moved away from the “form-only” proof-of-concept grounding by using Zhao et al.’s (2017) off-the-shelf pyramid scene parsing network (PSPNet), a state-of-the-art deep convolutional neural network for object segmentation in conjunction with a feedforward neural network for representing *object co-occurrences* (chapter 4). Our comparisons of the grounded semantic representations (*perceptual scene vectors* - PSVs) with conventional distributional language-based representations (both *latent* and *surface* semantic analysis) reveal the efficacy of using naturalistic scenes for grounding the meaning of concrete concepts. We also show that language surface structures encode meaning best when sufficiently constrained by modeller-determined feature sets, with performance deteriorating for randomly selected language surface structures. Furthermore, the meaning encoding of Latent Semantic Analysis improves as weaker dimensions are removed. These findings collectively indicate that although language is important, increasing the relevance of linguistic, statistical regularities is also critical. PSVs can semi-automatically extract strong associative and taxonomic relationships, measured both qualitatively and quantitatively. Critically, PSVs encode meaning without modellers hand-coding relevant features, which provides an ecologically valid approach to extending symbol interdependency beyond language and partially solving the *relevance problem* in semantics by grounding meaning extraction in real-world visual scenes. The statistical regularities in PSVs are sufficiently rich for meaning representation.

In chapter 5, we first replicate the *concreteness continuum* by re-analysing data from a large-scale normative study (Brysbaert et al., 2014). This is followed by extending PSVs using emotional expressions extracted from images (scene2vec) and demonstrating that grounded semantics leads

to high-quality representations for more concrete and some intermediate concepts while being inadequate for more abstract concepts like *freedom*. However, emotion-related inputs increase the quality of semantic representations, particularly for more abstract concepts. Our original contribution of modelling semantics using emotions only partially supports the *embodied abstract semantics* hypothesis (Kousta et al., 2011) and indicates that there is more to representing abstract meaning than emotions alone.

In the large-scale human semantics study on the geometrical properties and relations of meaning in chapter 6, we synthesise and extend Binder et al.'s (2016) research on brain-based componential semantic representations, and Troche et al.'s (2017) cognitive dimensions. Using our brain-based cognitive dimensions, we find that the local and global topological properties of the semantic network are best captured using non-linear dimensionality reduction (tSNE). We reveal that cognitive semantic networks have small-world properties and context-free semantic networks are organised lexically on a concreteness gradient. We also establish *scenes* as the most important semantic dimension, supporting a grounded perspective. Critically, the network topology of meaning is highly context-dependent, which lends support to our present thesis of grounding semantics in the real world, and that there is *no meaning without context*.

Lastly, in chapter 7, we extend our grounded scene2vec semantic representation using language-based tags, given the mutually-reinforcing nature of grounded and distributed representations (chapter 3) as well as the limits of grounding for representing more abstract concepts (chapter 5). Using two distinct image sources (*Google Images* and *Getty Images*) for training separate semantic networks, we find support for context-specific human-like gender biases. Our grounded semantic models can, therefore, represent well-established psychological traits, a prerequisite for grounded models to inform the mechanics of human semantic cognition. Using *semantic feature neutralisation* (SFN), we can selectively target and remove undesirable biases. In this final closing chapter, we aim to discuss

prominent critiques, which lead to us outlining future commercial applications of scene-based grounded semantics and future research directions.

8.4 Responding to Critiques of Grounded Cognition

8.4.1 Grounded Semantics and Dreyfus' Critiques

Hubert Dreyfus has been one of the staunchest critiques of AI (see Dreyfus, 1997, 2007), which leads to many in linguistics and psychology to use his arguments as support for a more embodied perspective on human cognition. Dreyfus' criticisms targeted the symbolic and logic-intensive GOFAI, to which even early rudimentary connectionist models from the 1980s would be strong rebuttals. However, somewhat surprisingly, Dreyfus (2007) more recently also attacked "embodied AI", which we interpret from the broader grounded AI perspective. The main issue is how embodied AIs will "directly pick up significance and improve our sensitivity to relevance since this ability depends on our responding to what is significant for us" (Froese & Ziemke, 2009, p.470). Our focus will be on Dreyfus' (2007) alleged failures of embodied or grounded AI as opposed to his earlier ones targeting GOFAI. In chapter 2, we outline and label (C1 to C5) Dreyfus' main critiques of grounded AI models, which we have inadvertently addressed in our present computational and empirical studies.

Dreyfus' overarching critique (C1) is that grounded intelligence (Brooks-style AI) does not have *situational awareness* - the ability to acknowledge relevant features under real-world conditions based on the system determining these particular features to be salient as opposed to being pre-determined by a human modeller (Froese & Ziemke, 2009). In our empirical research (chapter 6), we establish that meaning is indeed sensitive to *context-dependent relevancies* and that the relevant features (the cognitive

dimensions) are either stronger or weaker across different situations, such as *luminance* and *upper limb* dimensions being respectively dominant in the *house on fire* and *home move* scenarios.

From a computational perspective, our context-specific semantic networks reveal substantial differences in the quantitative and qualitative nature of meaning as a direct function of the training set (*Google Images* vs *Getty Images*). Therefore, the grounded semantic computational models demonstrate context-dependent relevancies. However, we concede, that all situational conditions across our grounded research are categorical and highly distinct (i.e. *house on fire* vs *gift giving*). Perhaps these large situational differences exaggerated the differences in the resulting semantic networks. Given the exploratory nature of much of our research, we opted to test broad general hypotheses. However, future research could test specific hypotheses investigating computationally-derived grounded semantic networks where the situation changes more continuously. For example, although we used naturalistic photographs for most of our computational studies, future research could use video footage and apply scene2vec to every *nth* frame, and develop a time-series based semantic network topology. This approach could also be applied to the study of routine action control, which has been based on either hand-engineered rule-bases (Cooper & Shallice, 2000) or hand-coded datasets (Botvinick & Plaut, 2002).

The second critique from Dreyfus (2007) claims that grounded AI lacks *understanding*. If confined to Brooks-style robotics or even more recent research on developmental robotics (Weng, 2004; Schmidhuber, 2006; Cangelosi & Schlesinger, 2015) Dreyfus' critique would remain unchallenged. However, our present work on grounded semantics is a first small step toward allowing AIs to develop an understanding of their environment. Our grounded computational models encode meaning that is not pre-determined by the modeller and is based on the environment and yet is comparable to language-based distributed models. This thesis goes

further. Not only do we claim that the environment needs to be understood by AIs, but that understanding itself is a grounded phenomenon. We do not claim that scene2vec representations have bona fide understanding, but we do maintain that grounded semantics is likely to be necessary for developing general understanding. Grounded cognitive models of semantics are likely to also bridge our theories of semantic memory in human and non-human primates and other animals. Seeking to account for semantics beyond language- or feature-based accounts is a necessary means for understanding the putative mechanisms of associative intelligence more broadly. Also, from an evolutionary perspective, the association cortex of the human brain does disproportionately expand compared to other higher primates (Buckner & Krienen, 2013).

Future work to critically evaluate, and perhaps, extend our present work might lead to Dreyfus' third criticism (C3) being addressed, namely the "inadequacy of current embodied AI for advancing our scientific understanding of natural cognition" (Froese & Ziemke, 2009, p.467). In our chapter 2 introduction, we argued that the inability for current "grounded" computational models of semantics to provide sufficient insights into human cognition is directly related to the GOFAI practice of modellers determining a phenomenon of interest and then hand-coding so-called "grounded representations" (e.g. Hoffman et al., 2018). Our study of the human-like biases found in scene2vec representations suggests that there might indeed be parallels between human cognition and grounded semantic models developed in the present thesis. Moreover, our proof-of-concept chapter 3 study on feature-based, grounded and hybrid representations support the earlier findings of Goldstone and Rogosky (2002), whose computational model discredited Fodor's (1998) influential theory claiming relations between concepts in a semantic system are insufficient for mapping concept correspondences. Computational models, grounded or otherwise, are capable of furthering our understanding of natural cognition, contrary to Dreyfus' assertion.

Another criticism, although more minor, of behaviour-based robotics, is that embodiment does not overcome the grounding problem (C4; Froese & Ziemke, 2009, p.467). Given that this thesis does not focus on embodiment but instead grounding, we suggest that our grounded semantic models are based on the dynamics of the real world. However, we propose that an exciting avenue for future research could be to investigate the interactions between embodiment, grounding and language, as their relative importance is likely to vary in a context-sensitive manner. Future work would ideally explore interdependencies between different formats of grounded associations, shaping the building blocks of higher-order cognitive semantics. One potential opportunity might be to examine scene-based correlates of the cognitive dimensions used in chapter 6's empirical investigation of semantic topology. This would require an extension of our computational implementation of grounding semantics, which is exclusively limited to vision.

Grounding a computational model or robot with the objective of having human-like experiences is unrealistic if we fail to provide a sufficiently realistic environment. Our use of naturalistic web-based photographs as opposed to dynamic live feed via cameras was a simplification due to computational processing limits. Though, future research could use parallelised GPUs or even custom-built TPUs (tensor processing units) for executing the most computationally demanding processing stages.

Dreyfus's (2007) fifth and substantial criticism of embodied robotics and AI is the *frame problem* (C5). The familiar "looser philosophical formulation" of the frame problem is the *relevance problem* - understanding what is relevant in a given circumstance. One of the greatest strengths of our scene-based grounded semantics is the non-division between meaning and context. This leads to a tight coupling between conceptualisations and perception. Meaning emerges from the continuous and discontinuous shifts in the frame itself. In scene2vec, objects that co-occur more frequently

across different images are also conceptually bound together. Here, the distinct images can be interpreted as frames. Therefore, grounded semantics overcomes the relevance problem through context-specific and bottom-up meaning encoding and representation. However, we acknowledge that Dennett's (2006) original interpretation of the frame problem, referring to first-order logic is not addressed by merely grounding associations in the real world. Nonetheless, we have shown that grounded semantics has already or at least has the potential for overcoming the five main critiques of grounded AI outlined in Dreyfus (2007).

8.4.2 Grounded Semantics and Johnson-Laird et al.'s Critique

Johnson-Laird, Herrmann and Chaffin (1984) provided a rigorous critique of traditional feature-based semantic networks like Collins and Quillian (1969) primarily as a result of connections only existing between concepts as opposed to with concepts and the world. Except for language-based semantic models, this critique from over three decades ago still applies to most cognitive models of semantics. Like Collins and Quillian's (1969) network model of semantic cognition, Anderson and Bower's (1973) *Human Associative Memory* (HAM) also consists of abstract propositional information, without reference to perceptually grounded information. These criticisms could be equally applied to even the most recent "embodied" variants of computational models of semantic cognition (e.g. Hoffman et al., 2018) where the propositional information is merely offloaded to artificially generated stimuli based on pre-determined similarities.

Our grounded model of semantic cognition, *scene2vec*, addresses Johnson-Laird et al.'s objection to only "concept-to-concept links" by breaking free from the symbol merry-go-round. We present a novel method for grounding semantics, in which concepts are rooted in both a number of naturally-occurring features in the real world and the resulting symbol interdependency between concepts themselves.

8.5 Future Applications

Despite a relatively long tradition of modelling semantics in psychology, to the best of our knowledge, there have not been any significant applied contributions to commercial applications using artificial intelligence (AI). Similarly, although cognitive science and neuroscience influence cutting-edge AI research (e.g. Silver et al., 2016), much of this is based on empirical research defining the boundaries of fruitful AI frontiers as opposed to implementing cognitive models. This is probably because traditional cognitive models of semantics are *hand-engineered, small-scale, effortful to build, and lack real information* given the *a priori* modeller specifications. Therefore, such cognitive models have limited, if any, applied utility in real-world applications. On the other hand, although not a cognitive model per se, distributed language-based semantic models like LSA and GloVe have been immensely successful in a wide range of business applications because of the scale of concepts represented, grounded in real human discourses.

Similarly, we suggest that this thesis' grounded *scene2vec* cognitive semantic representation will be more relevant for applied use cases in AI and robotics. However, we claim that visually grounded models ought to be especially suitable for modelling concrete real-world concepts. Moreover, more holistically grounded models (i.e. incorporating language and sensorimotor signals) are likely to lead to advancements in artificial general intelligence (AGI), transcending contemporary successes in narrow domains.

8.5.1 Implications for Artificial Intelligence and Robotics

Semantics is typically seen as the “holy grail” of cognitive science, semiotics, philosophy and neuroscience (Jackendoff, 2002), but also of artificial intelligence (Kiela et al., 2016). To this list of fields, we add robotics, given that sophisticated robots co-habiting a world with humans

would benefit from having a human-like understanding. To achieve this feat, we propose, robots' semantic representations ought to be grounded, at least partially, in the real world. However, would a "software-only" or virtual world suffice? Yes, but only if this virtual reality is sufficiently phenomenologically aligned with real-world properties. In other words, a GOFAI-style *blocks world* would be too simplistic to contain sufficient statistical regularities between objects within given scenes for an AI to acquire human-like semantic topologies. However, a high-fidelity virtual world like the *Real Sim City*²⁰ environment (see *figure 8.1*) that is based on real-world settings would contain most of the critical regularities necessary for grounded semantics to support the development of human-like semantic topologies in AIs and robots.



Figure 8.1: Depiction of two virtual scenes in *Real Sim City* modelled on real-world locations.

Here, we propose the novel idea of *virtual grounding* as an alternative to traditional hand-coded worlds but also the recently coined term *virtual embodiment* (Kiela et al., 2016). This is similar to using physics engines to acquire physics knowledge (e.g. Tassa et al., 2018). From our empirical research from chapter 6, we identified *scenes* as the most critical cognitive dimension across a broad range of concepts, while body-related

²⁰ <https://realsim.ie/realsim-city/>

dimensions (e.g. *upper limbs*) were less discriminating. This suggests that the environment we live in and the particular spatiotemporal properties of our experiences are likely to be more central for representing meaning than body-related regularities. A clear implication of this is that the AI or robot does not need a human-shaped body, but rather, needs to be placed in a human-like environment, for human-like meaning spaces to emerge. This argument challenges a wide range of embodied AI and robotics viewpoints becoming increasingly dominant over the last decade (Mainzer, 2009). We suggest that physically embodied AIs/robots (e.g. iCub) and virtual embodiment (virtual iCub) are less critical to semantics compared with the virtual grounding of AIs/robots. The advantages of virtual grounding, however, are comparable to those outlined for virtual embodiment, such as scalability, long-term feasibility (due to cost/effort), rapid iteration and largely human-free execution (see Kiela et al., 2016, for details).

Second, grounded semantic models are based on a core set of object and other scene-relevant regularities, which are likely to be reducible to a core number of dimensions, although these are likely to range from a dozen or so to several hundred as opposed to 3-dimensions, as suggested by Troche et al. (2017). These dimensions might be an opportunity for AI models, and robots relying on grounded semantic topologies to acquire context- and task-specific meanings but apply them more generally across spatiotemporal boundaries and novel situations never previously encountered. Thus, we suggest that grounded models of semantics are likely to further the study of *transfer learning* in contemporary artificial intelligence. In transfer learning, the knowledge represented based on a specific set of tasks and contexts can be applied more broadly to other tasks and contexts not previously encountered by the AI. This has significant implications for a plethora of AI applications. For example, contemporary state-of-the-art AI agents like *AlphaGo* (Silver et al., 2016) can beat world-class level *Go* champions but fail to beat a novice at chess, because of overly specialised intelligence, that fails to generalise to high-order game

strategies. Similarly, in the domain of meaning, text-based topic models can be trained on millions of snippets of text data on restaurant food reviews, but perform at chance-levels when classifying topics about movie reviews. Successful AIs that can navigate the complex tapestry of meaning require transferable meaning topologies to function across different tasks and scenarios intelligently.

Third, AI models and robots that have grounded semantic topologies and can bootstrap a set of common underlying scene-based dimensions are also more likely to develop *meta-learning* capabilities. Acquiring first-order semantic associations is a fundamental prerequisite for more complex behaviours to emerge. However, sophisticated logical and analogical semantic reasoning systems capable of solving real-world tasks need to make higher-order inferences.

One implication of our research is that this ever-increasing semantic abstraction can be grounded in cognitive dimensions. AIs and robots could be placed in increasingly sophisticated environments to ensure a gradual acquisition of semantic structures required for highly complex behaviours. Although some developmental roboticists (e.g. Georgeon & Cordier, 2014; Froese & Ziemke, 2009) advocate for training simple models and then gradually increasing complexity, this is typically achieved through different hand-coded levels. However, we, on the other hand, speculate that an AI or robot could develop across a graded series of virtually grounded environments that vary in complexity. At each *grounded level*, the AI not only learns the relevant first-order associations between objects but also higher-order relations such as the temporal dynamics of events, which could lead to a bottom-up grounded development of causality and perhaps even agency²¹. This speculation would need to be empirically tested.

²¹ Other mechanisms would also be necessary at this stage such as a forward model of an AI's actions.

Fourth, we propose, that AIs which rely on grounded semantic models are also more likely to display superior human-like *common sense* reasoning - a highly challenging task. There are numerous crowdsourced common sense relational datasets like *ConceptNet*, which contain triples such as “*pen*”, “*UsedFor*”, “*writing*” (Liu & Singh, 2004). However, constructing, maintaining and scaling these datasets is usually incredibly challenging and is reminiscent of the failed GOFAI attempts to codify “all” semantic knowledge in expert systems from the mid-1960s through to the early 1990s relying on *knowledge bases* and *inference engines*. Grounded semantic models are a possible candidate for implicitly representing these associations without the need for hand-coding particular relations. However, this is likely only to be fruitful if meta-learning capabilities are present since causal relations between objects will be a requirement. Future work on hierarchical learning might further clarify the limits and opportunities (if any) of grounded semantics representing more abstract concepts.

Lastly, the above implications lead to one of the most appealing and useful frontiers for grounded semantic models - developing *Explainable Artificial Intelligence* (XAI). Despite the vast plethora of successes in machine learning, much of contemporary achievements originate from relatively opaque deep learning models, see Shwartz-Ziv and Tishby (2017) for a detailed overview of why deep learning models are widely recognised as *black boxes*. This has a significant consequence on the usability of AI models and the deployment of robots in environments where humans roam freely.

The proliferation of XAIs is likely to be critical across a range of business, law, defence and medical domains because humans need to understand *why* an AI behaved in a particular manner, especially if human well-being is at stake. Traditionally, in the domain of machine learning, there is a trade-off between *performance* and *explainability*, with neural networks being high on performance (e.g. classification accuracy) but low on explainability, while decision trees are low on performance and high on

explainability (e.g. a list of *if-then* conditions). Current state-of-the-art techniques for XAI being developed by DARPA in a programme started in 2017 and aiming to finish in 2021 (Gunning, 2016), use *model layering*, where deep learning models and interpretable tree models are combined using *model induction*. Within this DARPA programme, decision rules can be extracted from the tree-based models inferring associations from the deeper explanatory neural network associations. The objective is to have AI models with high performance and explainability, although, we argue, this increases the overall model complexity.

On the other hand, our grounded semantics approach may not require sophisticated layering of different types of models because the underlying semantic dimensions are based on scene-based reinforcements. More complex grounded models of semantics with multiple levels of abstractions (e.g. causal inferences) could, in theory, be explained by exploring an AI's semantic topology's developmental pathways based on scene-level statistical regularities. This could provide an opportunity also to evaluate unwanted biases in the regularities learnt, which can then be debiased by selectively targeting and neutralising specific statistical associations. XAI is a growing sub-discipline within broader AI technologies and perspectives, but as AI solutions start becoming increasingly important and widespread, the onus will be on accountability through simple explanations. Grounded semantics might be one such candidate approach for XAI.

8.5.2 Implications for Advertising and FinTech

Here we explore potential applications of grounded semantics purely as a *monetisable technology* as opposed to a mechanistic cognitive model for furthering our understanding of how inter-concept dependencies can be rooted in our environment. Companies spend a highly variable proportion of their net revenues on annual advertising budgets. In the UK

alone, advertising spend is predicted to reach £20bn in 2019, a new record²², which excludes billions of additional spend on ancillary services ranging from creative industries to market research.

Traditionally, advertising is split into two categories, (i) *above-the-line* (ATL) mass media campaigns and (ii) *below-the-line* (BTL) personalised targeting via mail drops or emails. However, in the digital era of increased personalisation, advertising effectiveness has become a central priority for all Chief Marketing Officers (CMOs), ultimately responsible for advertising.

Companies typically have a *brand strategy* - a corporate view on the intangible meaning assets of a company to acquire new prospects and retain existing customers. For example, a traditional media conglomerate might want to differentiate itself from new digital native players by running nostalgic advertising campaigns combining emotional and rational messages about specific services through ATL and BTL campaigns. Conducting advertising research on a given set of campaigns can be prohibitively expensive and slow. Recruiting respondents, running a focus group or programming a quantitative survey with multimedia elements capturing the ad campaign is a complex activity requiring many skilled professionals. However, such research might indicate whether or not a given ad performs in-line with expectations. A problem with this format of research is that by the time the results are available, the campaign has already finished, which means the learnings from the research lack the tactical relevance of making any actionable recommendations or adjustments to the campaign. Therefore, all-too-often, campaign research "plays catch-up" by only helping to understand future advertising strategy as opposed to impacting tactical campaign execution.

²² Source: <https://www.thedrum.com/news/2018/06/20/uk-advertising-spend-track-top-20bn-2019>

However, with grounded representations such as *scene2vec* the video footage and imagery (e.g. for digital and print adverts) could be analysed without any primary research. The *scene2vec* outputs could help position the ads relative to other ads (akin to concept-to-concept relations) but also the underlying grounded *features* or *dimensions*. Relative ad positioning could help qualitatively and quantitatively provide timely feedback on whether or not a particular advert is strategically aligned with a company's desired brand associations²³. This could be done before the campaign goes live. For example, if a media owner wanted to have an emotional advert like the well-known *John Lewis* Christmas ads, then close alignment between the media ad and John Lewis' advert would suggest that to be the case. A company could compile a database of representative competitor ads and apply *scene2vec* on the various ad materials to extract semantic associations and output a network visualisation of all the ads in the database. Then, a new target advert could be evaluated using *scene2vec* to evaluate the relative position among its competitors. This would be an intuitive and easy-to-interpret method for evaluating ad-to-ad semantic associations.

Furthermore, these ads could also be explored not only based on the inter-ad comparability but also their underlying associations with the main features captured by *scene2vec*. For example, using our chapter 7 implementation (object co-occurrences, emotion expressions and linguistic tags), each of the ads could be evaluated against particular dimensions (e.g. *happiness*) for better understanding the underlying semantic cues. The combination of *ad-to-ad* and *ad-to-dimensions* diagnostics could help provide

²³ For example, the fast fashion brand H&M might be satisfied to have their ad be positioned near Zara, a slightly upmarket fast-fashion brand, but that would not be the case for a high-end fashion brand like Gucci or Hugo Boss.

quick and actionable tactical recommendations to help determine the execution format of the ad²⁴.

Similarly, in the nascent sector *FinTech*²⁵, machine learning algorithms analyse tens of thousands of data points for most consumers to deploy recommendation engines for personalising product offerings, and even customised prices. However, current technologies commonly use structured data and text-based unstructured data to make these recommendations. We propose that some of the digital image data could also be analysed using grounded scene2vec-type representations for understanding the *underlying meaning* associations from the websites already captured by the digital cookies²⁶. Each customer could have a digital *semantic* fingerprint capturing their unique lifestyle preferences.

Ideally, if the FinTech and banking sectors could be regulated to use these technologies responsibly, as opposed to exploiting customers by cross-selling unnecessary financial products, this could help banks develop superior *customer rewards* offers as opposed to the traditional catch-all set of services such as *cashback* and generic *retail vouchers* for shops and restaurants. In this scenario, a grounded semantic model could help the FinTech sector make useful personalised recommendations for banking products and services, based on data already being collected but not analysed. Banks could use this technology to enhance customer experience, reducing customer churn, and lower banking fees, which are typically inflated as a hedge for customer attrition.

²⁴ For example, the campaign execution duration of a suboptimal advert could be decreased to save money.

²⁵ A sector which uses “technology”, typically machine learning and block chain, to help the finance industry.

²⁶ This would not require any additional terms and conditions given GDPR regulations already apply to cookie data.

8.6 Future Research

In this penultimate section, we outline a selection of prospective semantic research avenues in the cognitive sciences, predominately focusing on cognitive modelling. The overarching theme throughout our subsections is the focus on using *real information* grounded in the environment as the building blocks of developing cognitive models of semantics, and ultimately, a mechanistically-grounded general theory of cognitive semantics.

8.6.1 Empirical Ground Truths for Semantic Modelling

There are currently no agreed benchmarks for comparing computational models of semantics with empirically-derived human meaning associations. Although this is less important for AI applications, it is fundamental to evaluating the efficacy of computational models in cognitive science. In this thesis, we have used several benchmarks, each with different strengths and weaknesses. First, we use language-based spaces such as LSA, GloVe and Skip-Gram for comparisons with our grounded representation. The main advantage of using these linguistic spaces for benchmarking cognitive models is the relative ease with which these spaces can be accessed. However, the limitation of using language-based representations is that they are not cognitive representations per se. Iteratively developing cognitive models to better capture statistical regularities solely found in language omits grounded information found in non-linguistic sources, like visual scenes.

Second, we use BrainBench, a small-scale neuroimaging-based representation, which has the advantage of representing different neural activation profiles across concepts, but is unlikely to capture all concept-related associations. However, we argue that scaling these types of datasets to human-level conceptual system scales will be prohibitively expensive making it impractical.

Third, using pure cognitive dimensions, without foundation in neuroimaging studies, like in Binder et al. (2016), is scalable but ultimately not as parsimonious given that it is difficult to generate an explicit set of criteria for inclusion of particular dimensions. Fourth, in chapter 6, our hybrid technique might be optimal for exploring scalable ratings across thousands of concepts as well as being based on neuroimaging foundations. This could also lead to future explorations of computational lesion studies to simulate particular neural atrophies in a grounded computational semantic model.

In our empirical study of conceptual topologies, we find evidence in favour of dynamic meaning structures, which are subject to context-based shifts, leading to our conclusion of there is *no meaning without context*. Therefore, static meaning spaces are unlikely to be particularly useful for understanding everyday mental representations of meaning outside the sterile conditions of laboratory experiments with decontextualised concepts. We suggest that future investigations use 10,000+ concepts as opposed to our circa 500 words but simultaneously conduct more nuanced smaller-scale contextualised experiments with dozens of words in real-life settings. Our experiment with ad hoc categories is an “extreme conditions” test (e.g. *house on fire* versus *cooking*) to see if topologies shift. Now, that we have established topological shifts, refined hypotheses regarding concept, task and context combinations can be investigated. This will change the focus to a dynamic study of meaning as opposed to the present-day static conceptualisations of semantics.

8.6.2 Large-scale Cognitive Semantic Modelling

We believe our scene-based grounded semantic models will scale more effectively as a result of not requiring hand-coded feature sets. Future cognitive modelling studies of semantics could start using even more ecologically valid stimuli than our present studies (e.g. *video streams*), which could incorporate other forms of grounded statistical regularities. Sound

signals from videos could be analysed using *natural language processing* (NLP) and *Mel Frequency Cepstral Coefficients*²⁷ (MFCC), to respectively process language (*communication dimension*) and sounds (*auditory dimension*). Building on previously discussed neuroimaging work (Binder et al., 2016), future studies could also incorporate dimensions beyond the five primary senses.

The cognitive dimension *luminance* could be measured by quantifying the brightness levels of particular objects in scenes. The dimensions *upper limb* or *ingestion* could be evaluated using pre-determined upper body *regions of interest* and a CNN such as *PoseNet*²⁸. The dimension *human* could be represented based on the presence/absence of humans in a particular video frame. This could ultimately lead to a computational model of semantics that is not only applied to dozens or even a hundred concepts as in chapter 5 of this study but tens of thousands of concepts, gradually approximating the size and complexity of the human semantic memory system. This has the potential to raise a plethora of new research questions, some of which might help further streamline our current pluralistic perspectives on representing meaning. For example, do small-world properties of grounded cognitive semantic networks change as a function of network size? Are grounded representations better at capturing semantic associations in larger systems? How do context-specific shifts in the topology manifest in the network at different scales? Although somewhat far-fetched, could grounded representations be used to help translate semantic spaces across different animals, if salient semantic dimensions are represented appropriately, e.g. for dogs, *olfactory features*?

²⁷ MFCCs are known to represent sounds in a similar manner to the human auditory system (see Kiela & Clark, 2017).

²⁸ PoseNet is deep learning based pose estimation model, which is part of the OpenCV toolkit.

We believe that some of our original analytical investigations of cognitive semantics, for example, network visualisation, analysis and regularisation techniques might also be particularly useful for interpreting meaning representations as systems scale and non-linear complexities compound at an increasing rate.

8.6.3 Development of Semantic Network Topologies

Developmental cognitive roboticists narrowly investigate grounding from the dual perspectives of *embodiment* and *language* but overlook the role of the environment in conceptual development. For example, Mirolli and Parisi (2009) adopt a Vygotskyan perspective on conceptual development in robots, based on increasing linguistic sophistication. Similarly, Lallée et al. (2010) propose a cognitive robotics framework based on linguistic and embodied regularities, while Farkaš, Malík and Rebrová (2012) use reinforcement learning on sensorimotor representations for grounding meaning. However, none of these studies explores the role of grounding meaning in real-world cues or develop topologies for investigating concept relations. Therefore, it follows, that these studies do not explore the gradual developmental shifts in computationally grounded semantic topologies, which could provide a mechanistic testbed for developmental theories of psychology.

Future computational studies might capture scene-based information typically encountered by infants, toddlers and young children through the use of body-mounted cameras. Grounded computational representations like scene2vec could then, over time, model conceptual topologies and evaluate changes in the structure of the semantic networks. This could be done for typical and atypical populations to investigate the role of the environment. For example, how might a physical disability or emotional distress in early childhood constrain exposure to grounded real-world associations? We predict that network properties such as small-worldness are likely to emerge at a particular stage when long-range links

between commonly unrelated concept node clusters start emerging. Such computational models could be evaluated against behavioural assessments of cognitive performance.

Grounding and investigating semantic networks is merely the first step. The present research has shown that even representing the topologies of smaller sets of concepts can be highly dynamic and context-specific. This is likely to increase in complexity as studies incorporate scene-based and language-based grounding in conjunction with embodied grounding using sensorimotor regularities. Future computational studies could “bring to life” these holistically grounded semantic topologies by incorporating spreading activation mechanisms to, for example, test these models using behavioural semantic priming associations and higher-order analogical reasoning tests.

8.7 Conclusion

The overarching objective of this thesis has been to further the field of grounded semantics using a range of computational and empirical studies. Our approach has predominantly used non-linguistic scene-based grounding of semantics as an alternative to both hand-coded features and embodied sensorimotor signals.

First, we resurface grounded cognition’s origins in AI and robotics which paved the way forward for linguists and cognitive scientists to investigate bodily and situational factors shaping meaning. Second, we show that hybrid representations comprised of feature-based and grounded stimuli are more robust than either format individually. Given that cognitive semantics has traditionally overlooked the grounded perspective, we make this our focus. Third, we extend *symbol interdependency* by revealing that language surface structures encode meaning particularly well when constrained by modeller-determined feature sets. Critically, we demonstrate the feasibility of semi-automatically

extracting strong associative and taxonomic relations from object co-occurrence relations in naturalistic images.

Fourth, we show that despite object co-occurrences being validated against a neuroimaging benchmark, grounded semantics is better at representing more concrete than abstract concepts. Combining object co-occurrence regularities with emotion expressions, *scene2vec*, however, improves the quality of abstract semantic representations. Fifth, in our empirical study, we reveal the small-world network topology of the human meaning space, which is structured lexically in a neutral context, while contextual shifts dynamically modulate this topology. This study also reveals the dominance of *scenes* in human semantic memory. Sixth, we find context-dependent human-like biases in our *scene2vec* representation, which supports the psychological plausibility of grounded representations.

In conclusion, this thesis has provided support for a novel computational viewpoint on investigating meaning - *scene-based grounded semantics*. Future research scaling scene-based semantic models to human-levels through *virtual grounding* has the potential to unearth new insights into the human mind and concurrently lead to advancements in artificial general intelligence by enabling robots, embodied or otherwise, to acquire and represent meaning directly from the environment.

Appendix A

Network Analysis

A.1 Overview

In chapters 6 and 7, several types of network analyses and visualisations are run for computationally and empirically investigating grounded semantics. We provide additional technical details and definitions related to networks in this Appendix.

A.2 Graph Theory

Graph theory is the mathematical study of networks dating back to the 18th century. Using a popular puzzle (*Koenigsberg bridge problem*), Euler (1736) proved that a path by which someone could cross all seven bridges exactly once and return to the starting point did not exist (see *figure A.1*). This finding demonstrated the advantages of abstracting distance and graphical position, which gave rise to the idea of *geometry of position* (*geometria situs*). This example also shows the real-world or applied origins of graph theory.

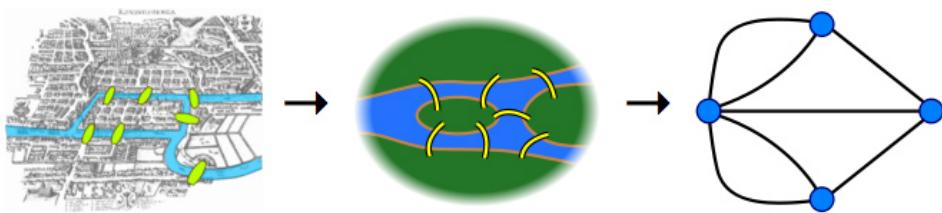


Figure A.1: An overview of the Königsberg bridge problem at different levels of abstraction. Source: <https://plus.maths.org/content/bridges-k-nigsberg>

A.3 Network Analysis

Network analysis is based on graph theory and is the science of characterising **networks**, an abstract system comprised of a set of **nodes** (elements of a system) and **edges** (links between elements). Edges can be **directed** or **undirected**, where the former indicates asymmetric information flow, while the latter, symmetrical information flow. In this thesis, we exclusively focus on undirected edges.

A network's most basic data representation is the **adjacency matrix**, which, in a **binary graph**, contains the presence (1) or absence (0) of a link between all node pairs (see *figure A.3*). This is the data structure generated from the associations (e.g. *correlation* or *distance*), by including strong neighbours through applying a threshold to the associations, which subsequently leads to visualising a sparse network.

The *physical layout* of the network is only important relative to the node and edge relations. The length of edges is not important, and nor is the position of nodes in 2D space. We use the **Fruchterman-Reingold**, a force-directed or spring algorithm, for determining the layout of network visualisations (Fruchterman & Reingold, 1991). This technique, as well as other layout variants (see *figure A.2*), do not alter the network structure (node-to-node relations) but only the format in which a network is depicted. Our motivation for using the Fruchterman-Reingold algorithm is two-fold. First, it yields network visualisations that are easier to interpret because of non-overlapping edges, and second, it is aesthetically pleasing.

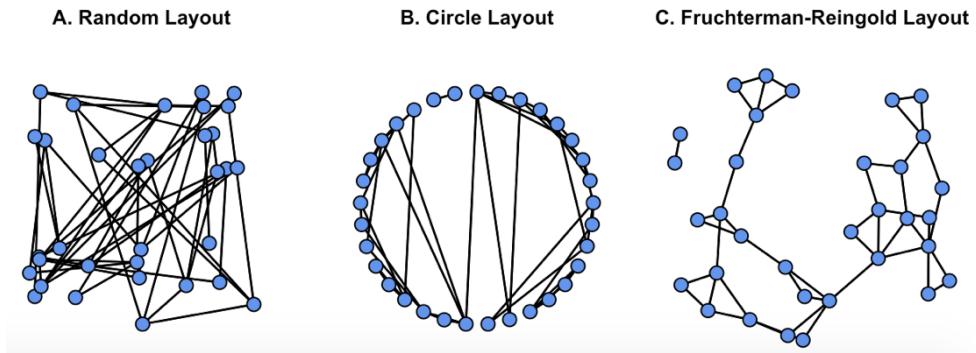


Figure A.2: The same network visualised using three layouts, which are *random* (A), *circle* (B) and *Fruchterman-Reingold* (C). The Fruchterman-Reingold layout has the advantage of greater interpretability.

A.4 Main Network Metrics

The **size** of the network is simply the number of nodes present. In **connected graphs**, there is only one component, see *figure A.3* for an example of a **connected graph**. Networks can also be split into several subgroups, where the number of subgroups determines the **components** metric. The **diameter** of a network is a helpful measure for describing compactness, based on network **paths** - the number of steps required to go from one node to another. In *figure A.3*, the **longest path** to get from node A to E is 4 (A-B-C-D-E), while the **shortest path** is 2 (A-D-E). The **diameter** of a network or its component is the “longest of the shortest paths across all pairs of nodes” (Sporns, 2011, p.15).

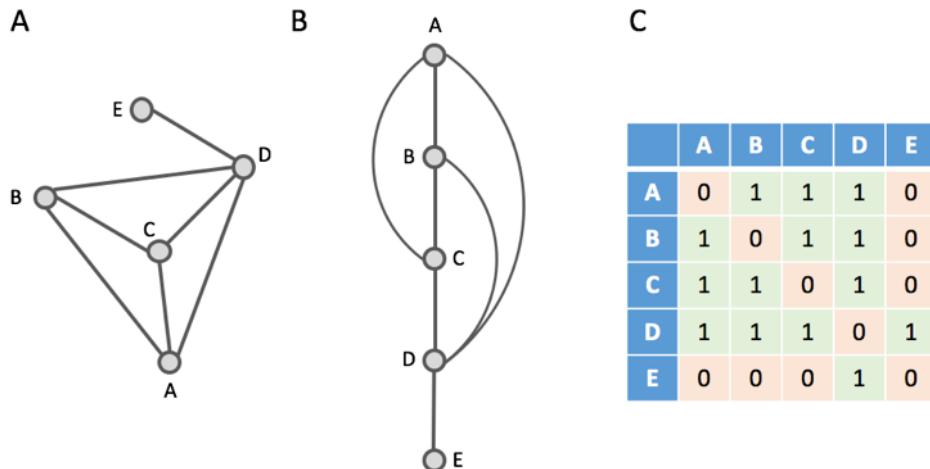


Figure A.3: The same network structured *geometrically* (A), a *cascaded system* (B) and an *adjacency matrix* (C), showing the presence (1) and absence (0) of edges between node pairs.

Finally, the **density** of the network is one of the main network measures and refers to the proportion of edges available in a given network to the maximum number of possible edges, ranging between 0 to 1. Directed and undirected graphs have different density calculations given that in directed graphs, edges between two nodes are counted twice, while only once in undirected graphs.

In the case of undirected graphs, the maximum number of possible edges among k nodes is $k * (k - 1)$, and the formula for density, shown below, also includes L , the number of observed edges. In other words, the density of a network is the number of *actual links* as a proportion of the number of *possible links*. In the network from *figure A.3*, the network has a total of 5 nodes (k), and 7 edges (L), resulting in a network density of 0.7. We illustrate how network density changes in a graph with four nodes and a varying degree of links in *figure A.4*.

$$\text{density} = \frac{2L}{k \times (k - 1)} = \frac{2 \times 7}{5 \times (5 - 1)} = 0.7$$

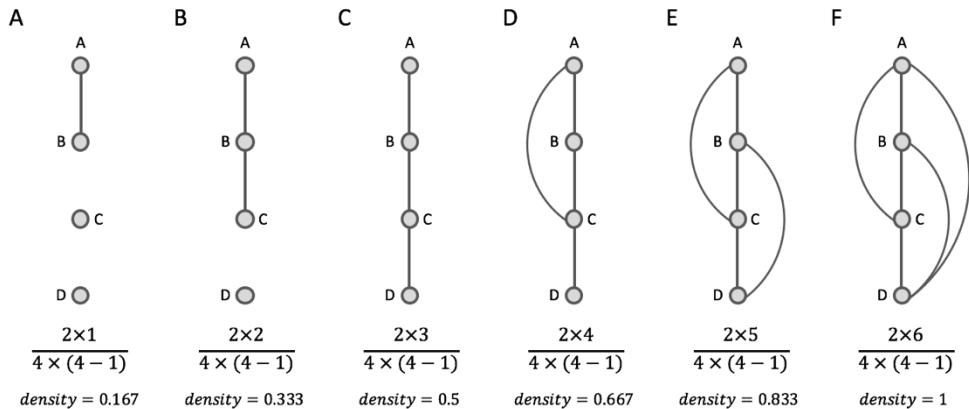


Figure A.4: Example of how network density changes as a function of increasing the number of links (L) from a single link (A) to all six links (F). The networks A-F all have four nodes (k).

Appendix B

Limits of Simple Plots

B.1 Overview

In chapters 6 we outlined the advantages of using non-linear visualisation techniques such as t-SNE (t-distributed Stochastic Neighbour Embedding) for higher-dimensional datasets. Here we briefly outline the main difficulties of using simple plots (e.g. radial graphs) to represent multidimensional datasets.

B.2 Challenges of Visualising Multiple Dimensions

In machine learning and data science more broadly, one typically uses *parallel plots*, *radial graphs*, or even *word clouds*. One of the critical constraints with all of these more traditional data visualisation techniques is that they each visualise a limited number of dimensions simultaneously. For example, word clouds typically represent word frequencies in the form of the size of the words, where the position of the words can be random or forced to fit a particular shape. So, if the word *calculator* occurs in a corpus

twice as frequently as the word *tablet*, then the former might be twice the size of the latter (or some other pre-set scaling factor).

Further elaborations of the word cloud (e.g. adding colours), could add a new dimension to the plot. Additional attempts at discriminating extra dimensions (e.g. changing the font of the words) can be hard to perceive unless a highly intuitive dimension is chosen to be represented by different fonts, such as “older items” being depicted by *old English* fonts while newer items in modern fonts like *Arial*.

However, visualising beyond two- or three-dimensions is rarely feasible as the additional dimensions layered on top of existing dimensions leads to *visual interference*. Thus, this type of layering of one dimension on top of another dimension rapidly reaches its maximum utility after three dimensions.

The reduction of dimensionality in a dataset can be achieved by applying traditional linear dimensionality reduction techniques. Popular linear variants are *principal component analysis* (PCA) and *multidimensional scaling* (MDS), while *t-SNE* is an increasingly prominent non-linear dimensionality reduction technique.

Appendix C

Limits of PCA and MDS

C.1 Overview

In chapters 3 and 4 we use multidimensional scaling (MDS), but in subsequent chapters with typically more concepts being modelled, we use t-SNE for dimensionality reduction. Here we briefly outline the challenges of using PCA and MDS for modelling semantics.

C.2 Dimensionality Reduction

In cognitive semantics research (e.g. Rogers & McClelland, 2004) as well as in our computational analyses in chapters 3 and 4, MDS can map higher-dimensional structures in lower-dimensional maps consisting of two or three dimensions. MDS can represent the concepts such that similar concepts are mapped more closely together while different ones further apart. Ideally, the distances in the lower-dimensional map reflect the similarities and dissimilarities of the higher-dimensional representation. For example, PCA finds a linear projection of higher dimensional data

points by maximising the variance of the projected data, which relies on the preservation of large distances (Jolliffe, 2011).

The classical example of demonstrating this idea in machine learning uses the so-called “Swiss-roll”, where the Euclidean distance is not the best measure of distance between two points in non-linear manifolds. In the case of the Swiss-roll, *geodesic distances*, as opposed to Euclidean distances, are more suitable. In *figure C.1* we show how a small Euclidean distance between two points is a large geodesic distance, where the latter is a better approximation of the non-linear manifold representation.

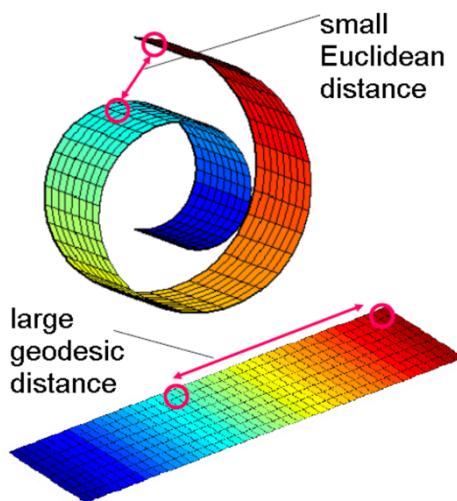


Figure C.1 Example of comparing Euclidean and geodesic distances on a 3-dimensional flat Swiss-roll. Source: <https://www.cs.cmu.edu/~epxing/Class/10715/lectures>.

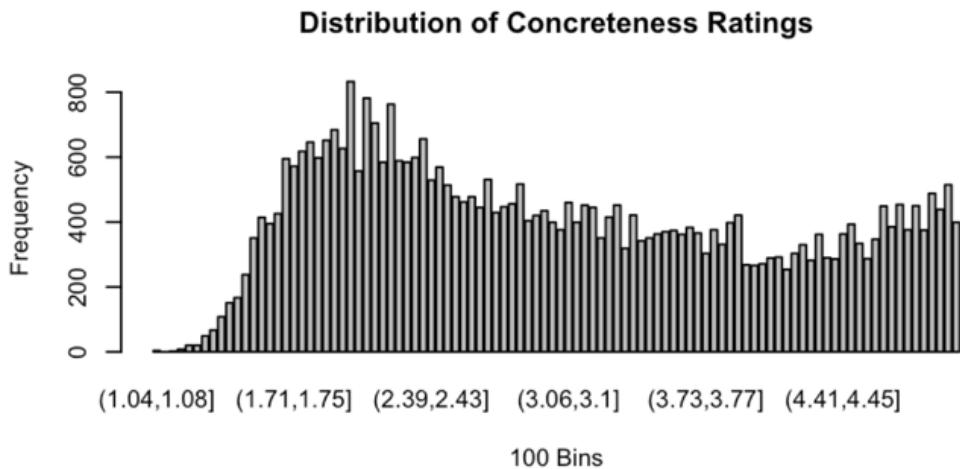
We argue that the cognitive semantic space is an example of highly complex and non-linear topology, and therefore not best suited to linear dimensionality reduction, which favours larger distances in space while discounting smaller ones. In the computer science literature (e.g. Gisbrecht, Schulz, & Hammer, 2015), the focus on smaller and larger pair-wise distances are, respectively, referred to as either preserving *local-* or *global-structures*. Thus, in traditional cognitive science visualisations of semantic spaces, we argue, the semantic space’s quality is compromised by not focusing on the local structures inherent in the high-dimensional input data due to optimising an objective function mathematically designed for maximally accounting for global structures.

Appendix D

Enlarged Images of Results

Throughout the thesis, some images are too small for detailed evaluations in the print version. Therefore, in this appendix we provide enlarged formats of a selection of our results where the readability of charts needs to be enhanced. This is done either by amending the page layout or by splitting the images into multiple panels across different pages. The figures appear in chronological order as presented in the thesis along with their figure annotations.

A



B

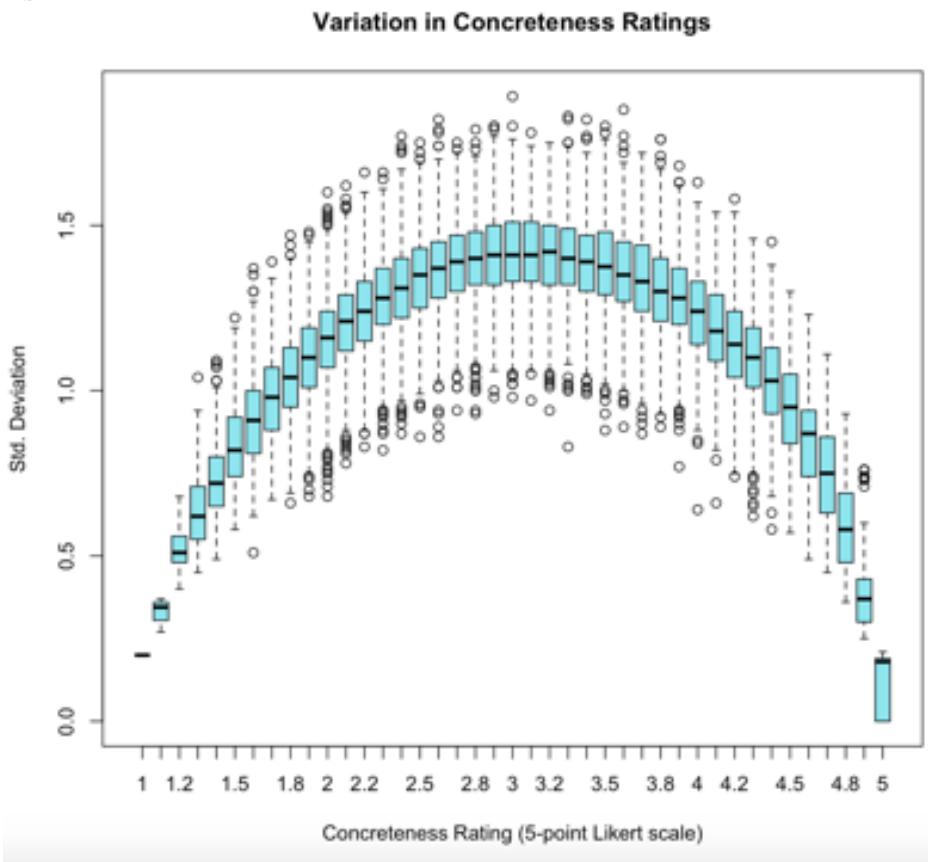


Figure 5.2 (A) Histogram of the concreteness ratings of 40,000 word lemmas (B) Boxplot of the standard deviations of the same concreteness ratings.

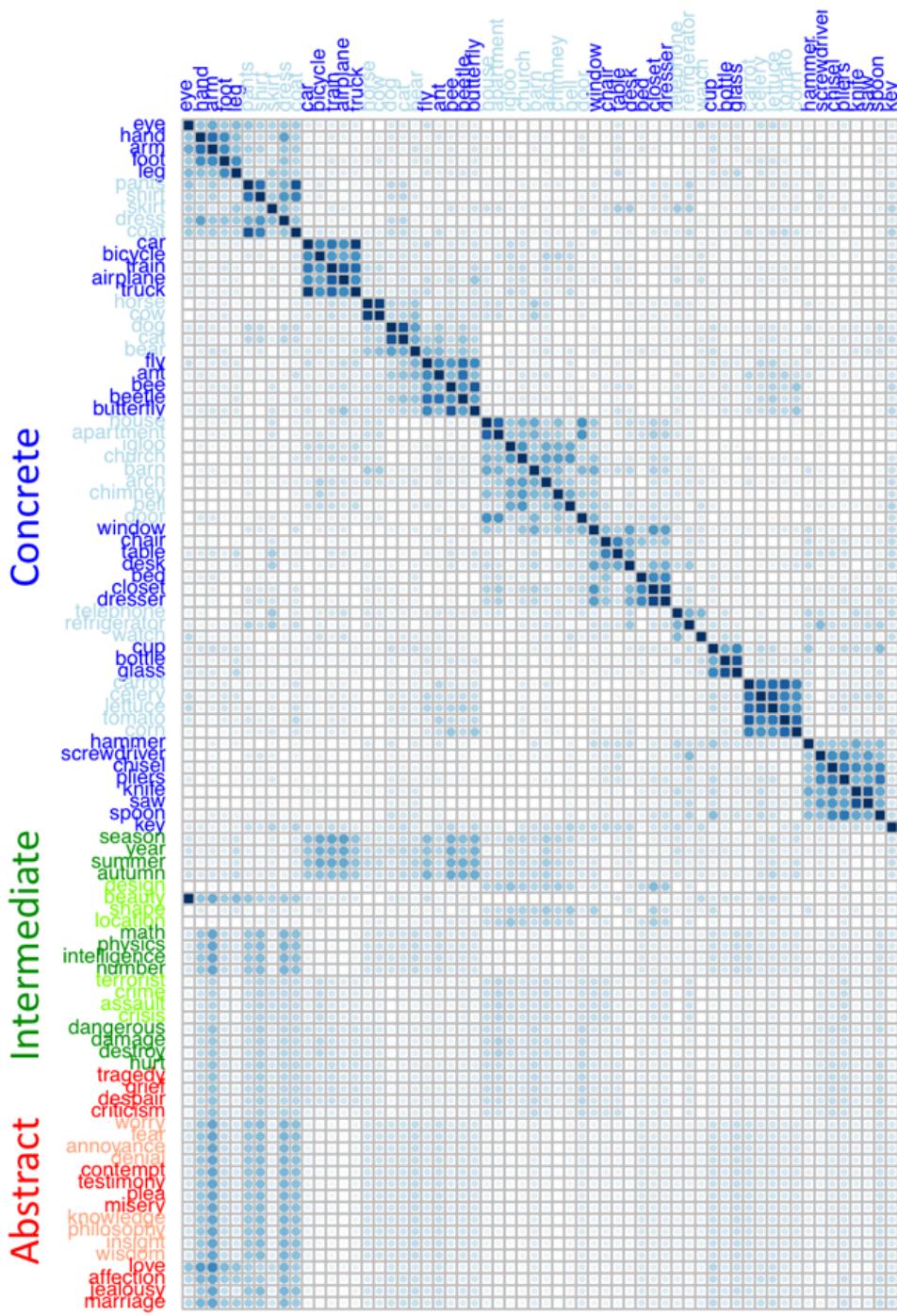


Figure 5.4: *Left Panel:* A correlation plot of the PSV's hidden layer representations. Concepts are grouped into concrete (blue), intermediate (green), and abstract (red) groupings. Within both the intermediate and abstract groups, concepts are grouped into the LSA-based “concept clusters”, and these are highlighted by respectively alternating between darker and lighter shaded of green and red. Similarly, we also alternate darker and lighter shades of blue for concrete words but use the original order of concepts used by Xu et al. (2016).

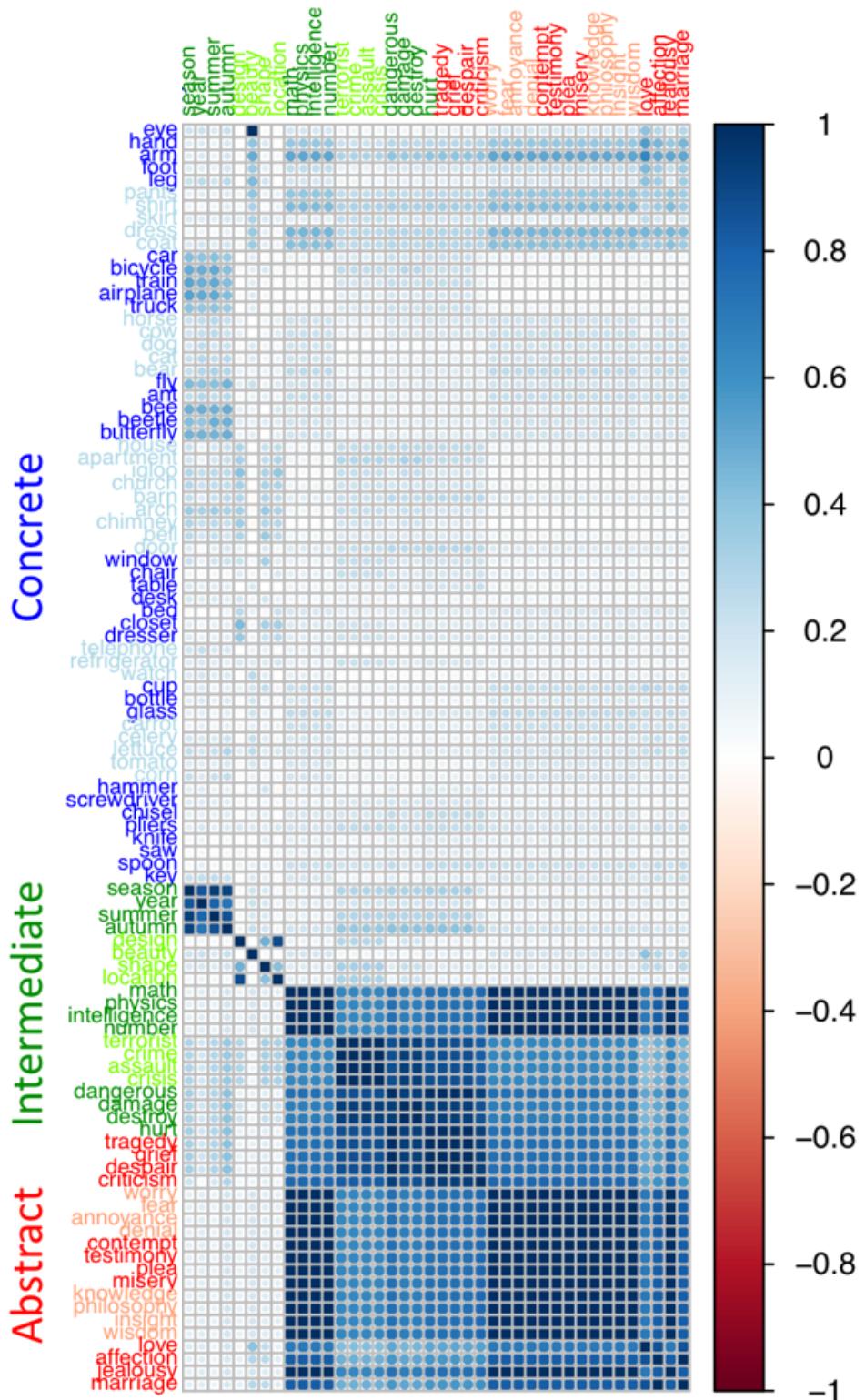


Figure 5.4: Right Panel: A correlation plot of the PSV's hidden layer representations. Concepts are grouped into concrete (blue), intermediate (green), and abstract (red) groupings. Within both the intermediate and abstract groups, concepts are grouped into the LSA-based "concept clusters", and these are highlighted by respectively alternating between darker and lighter shades of green and red. Similarly, we also alternate darker and lighter shades of blue for concrete words but use the original order of concepts used by Xu et al. (2016).

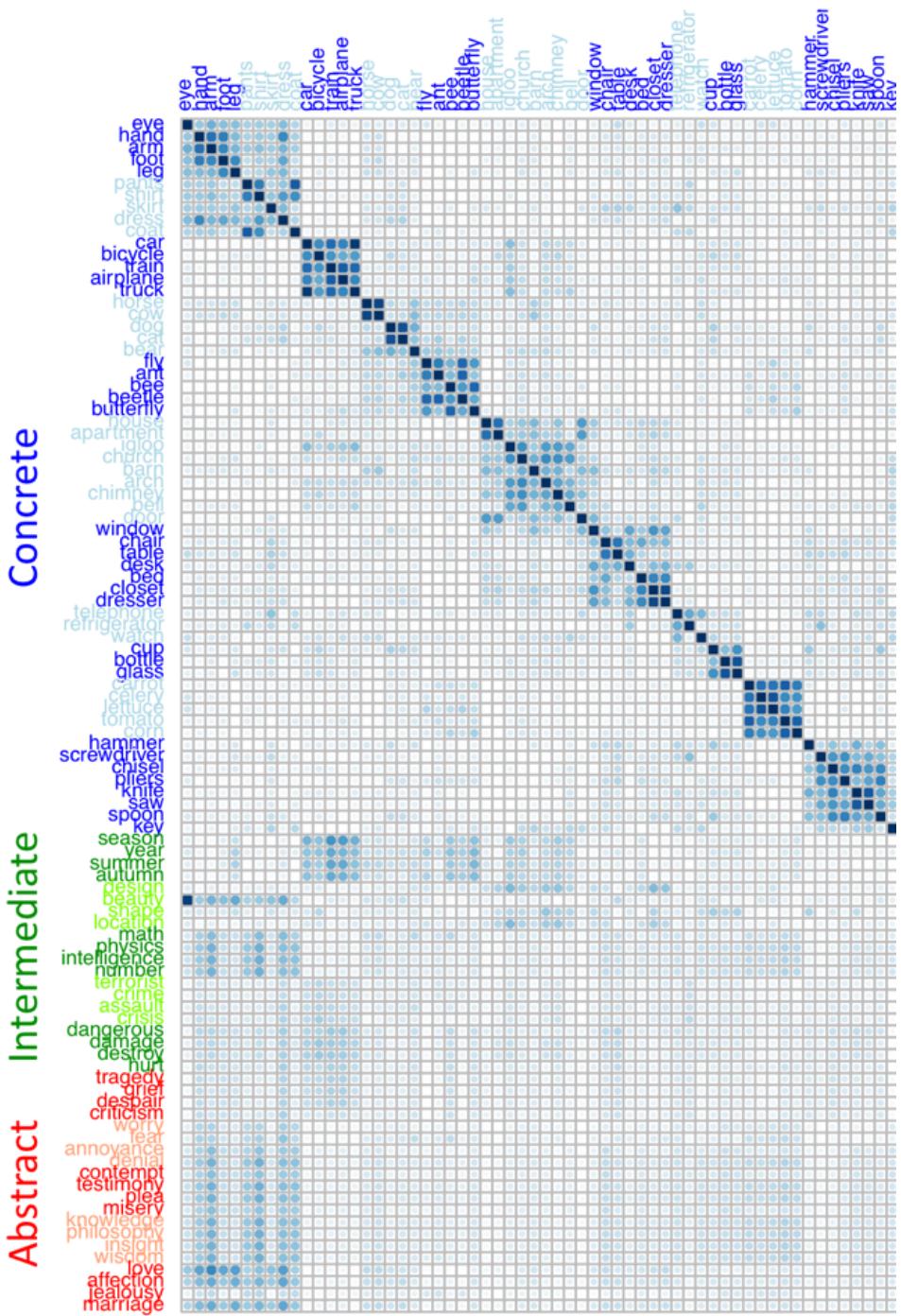


Figure 5.8: Left Panel: A correlation plot of scene2vec's hidden layer representations. Concepts are once more grouped into concrete (blue), intermediate (green), and abstract (red) groupings.

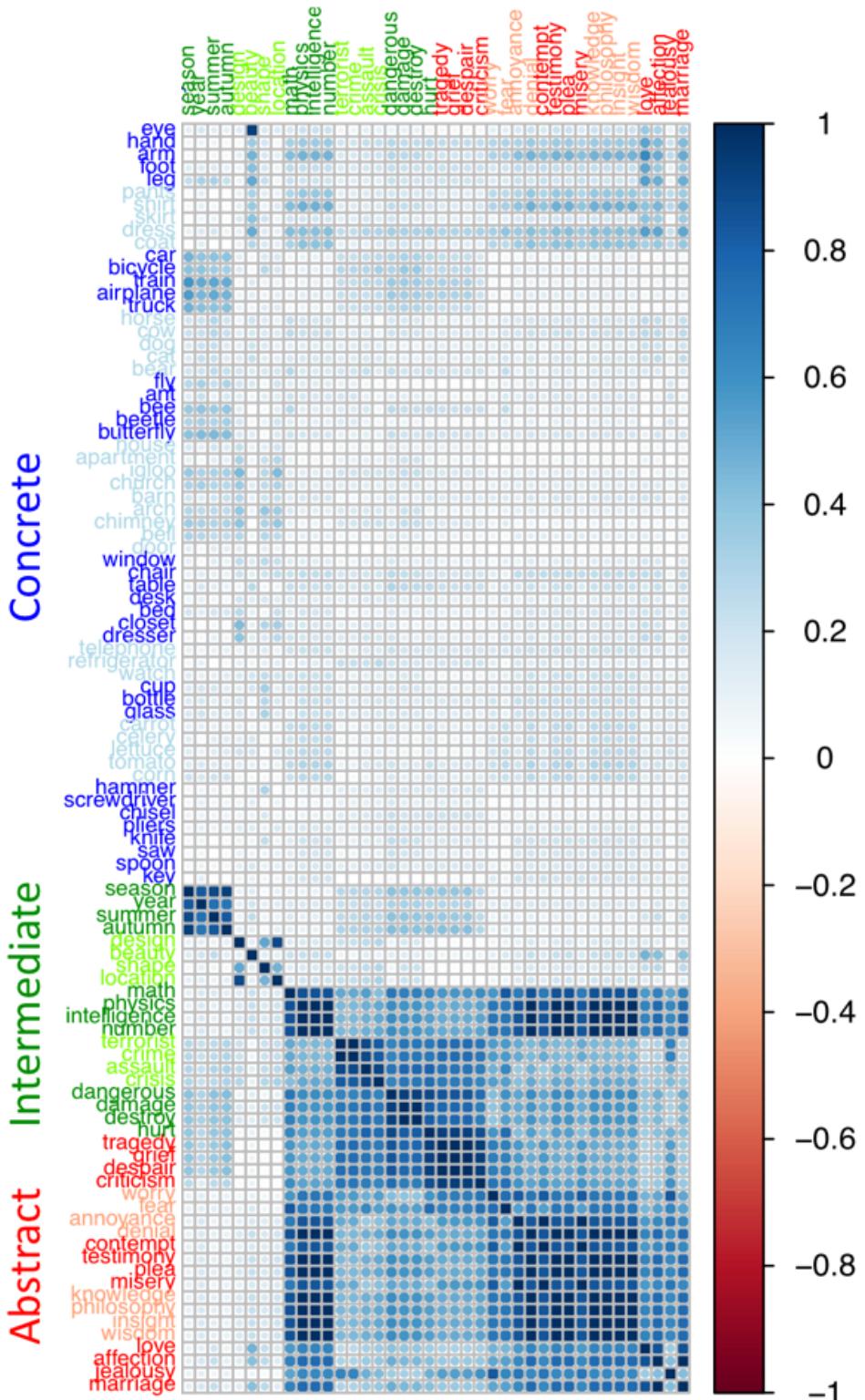


Figure 5.8: Right Panel: A correlation plot of scene2vec’s hidden layer representations. Concepts are once more grouped into concrete (blue), intermediate (green), and abstract (red) groupings.

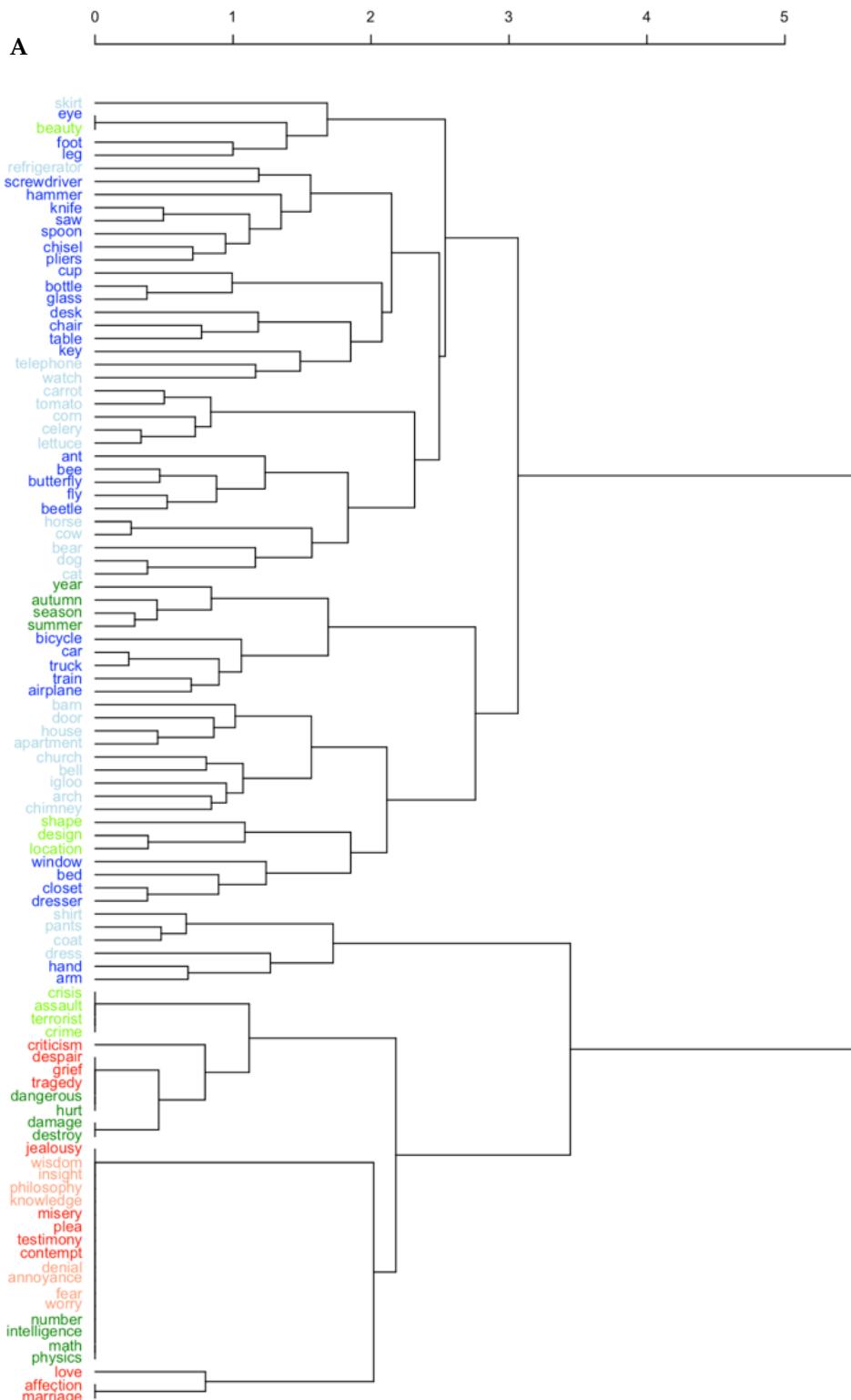


Figure 5.9A: Hierarchical cluster plot of the hidden layer neurons representing the semantic associations of PSVs. Concepts are once more grouped into concrete (blue), intermediate (green), and abstract (red) groupings.

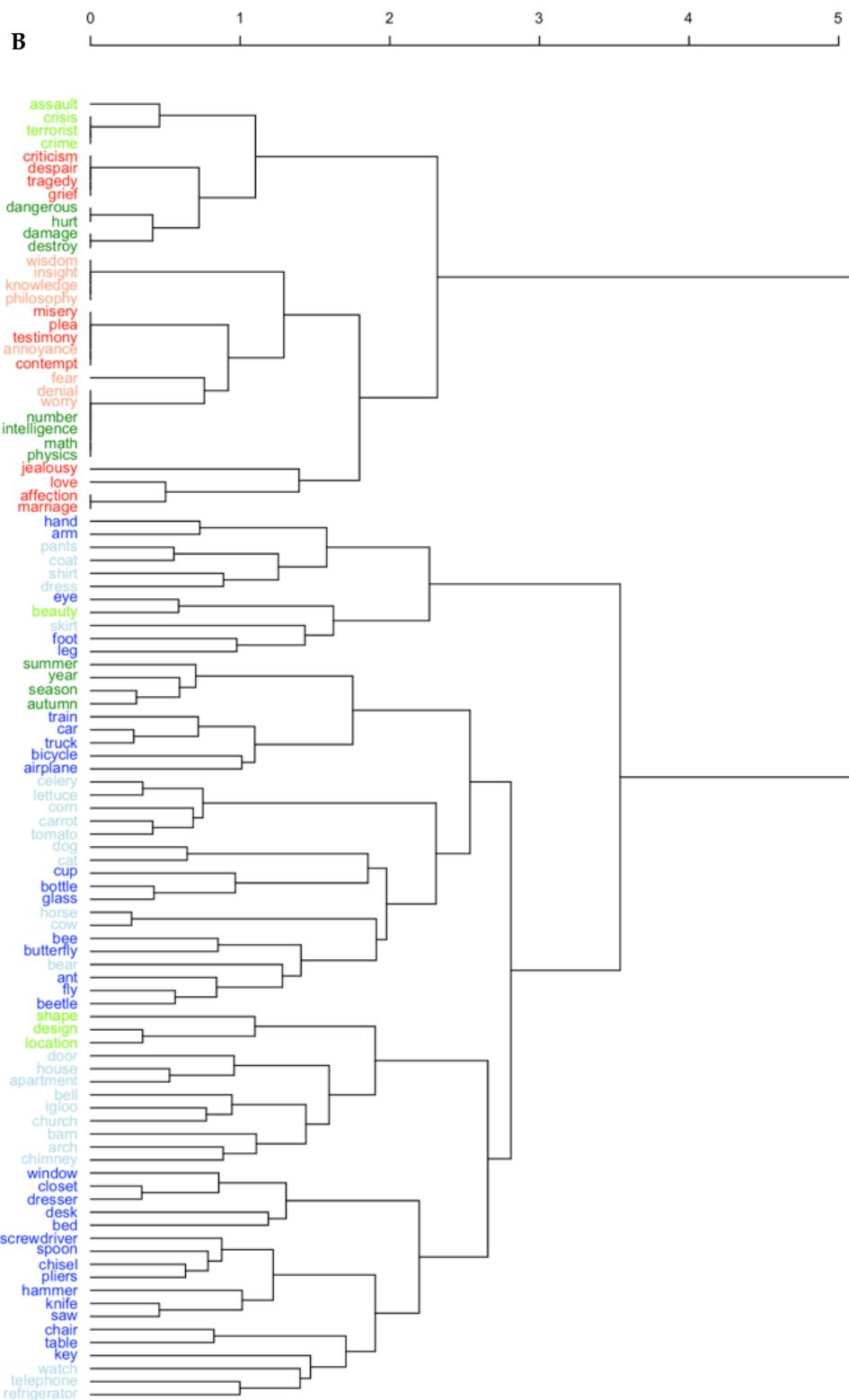


Figure 5.9B: Hierarchical cluster plot of the hidden layer neurons representing the semantic associations of scene2vec. Concepts are once more grouped into concrete (blue), intermediate (green), and abstract (red) groupings.

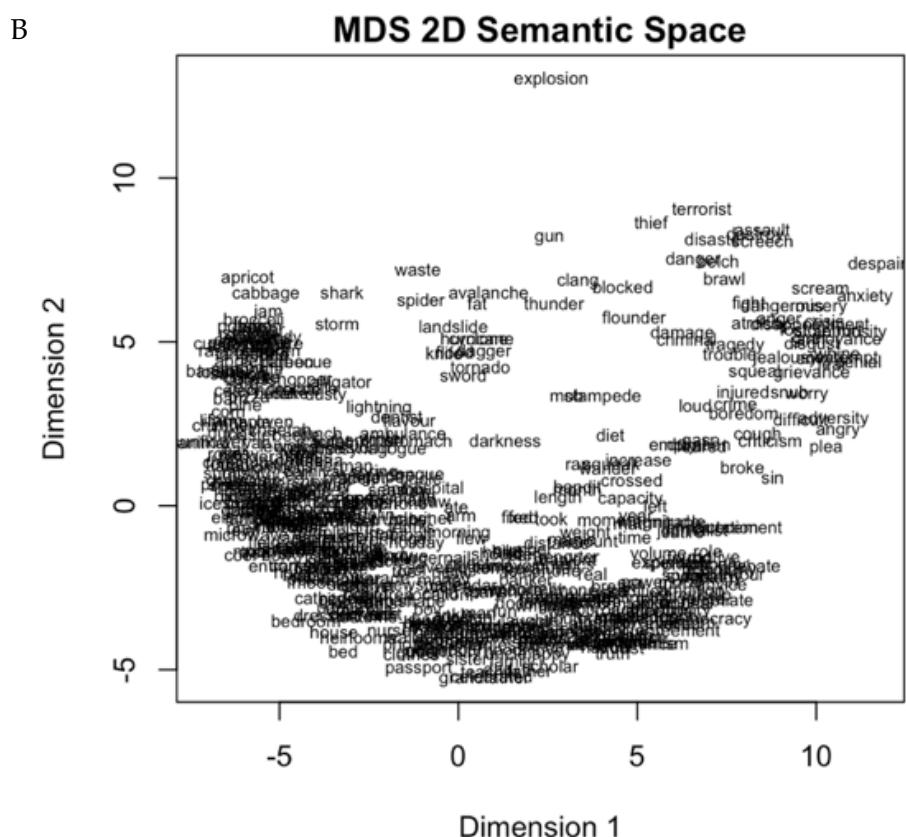
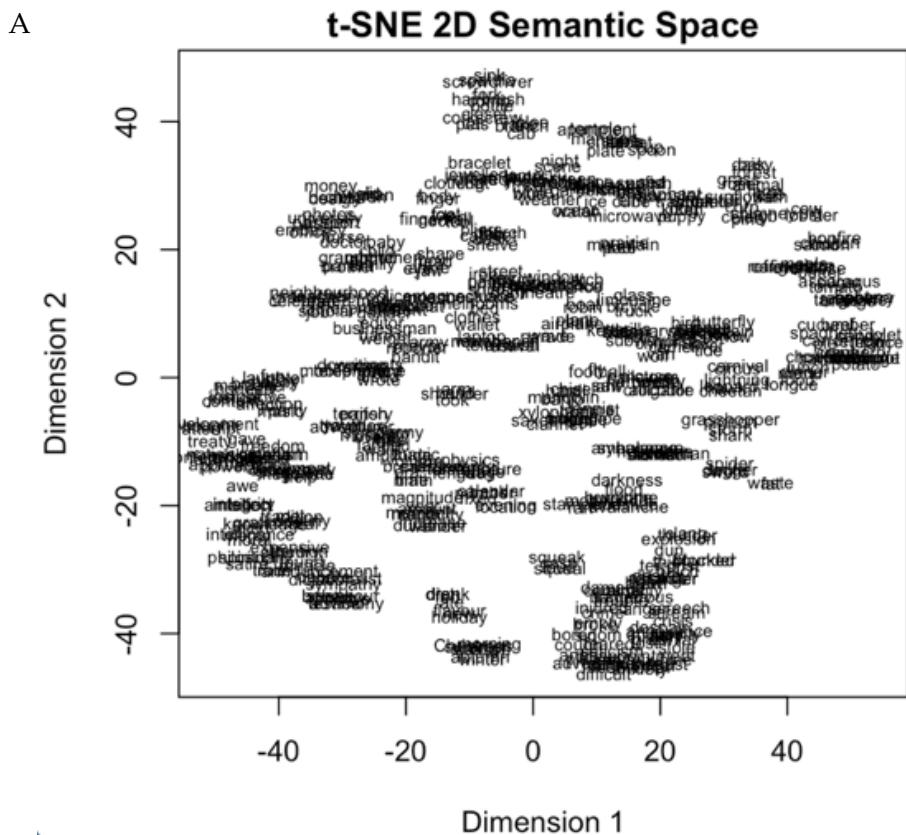


Figure 6.9: Comparing t-SNE's semantic embedding space (A) with MDS' space (B) for all 544 concepts.

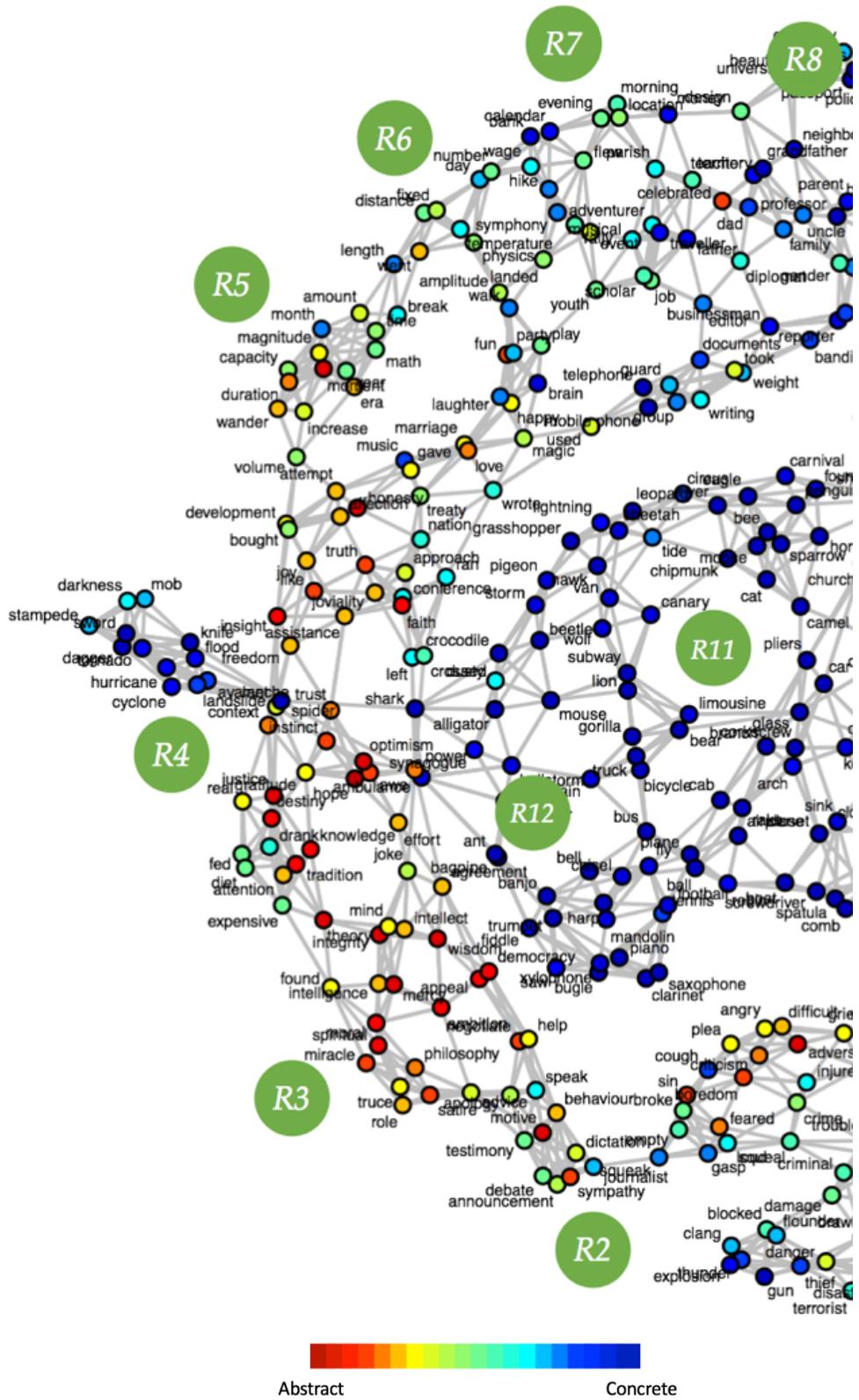


Figure 6.13: *Left Panel:* Overall semantic topology, including concept labels and the nodes of the network colour-coded according to the concreteness spectrum, ranging from **red** (abstract) to **green** (intermediate) and **blue** (concrete). The green numbered circles (i.e. R1...R14) highlight different portions of the network to aid discussion of more specific network neighbourhoods.

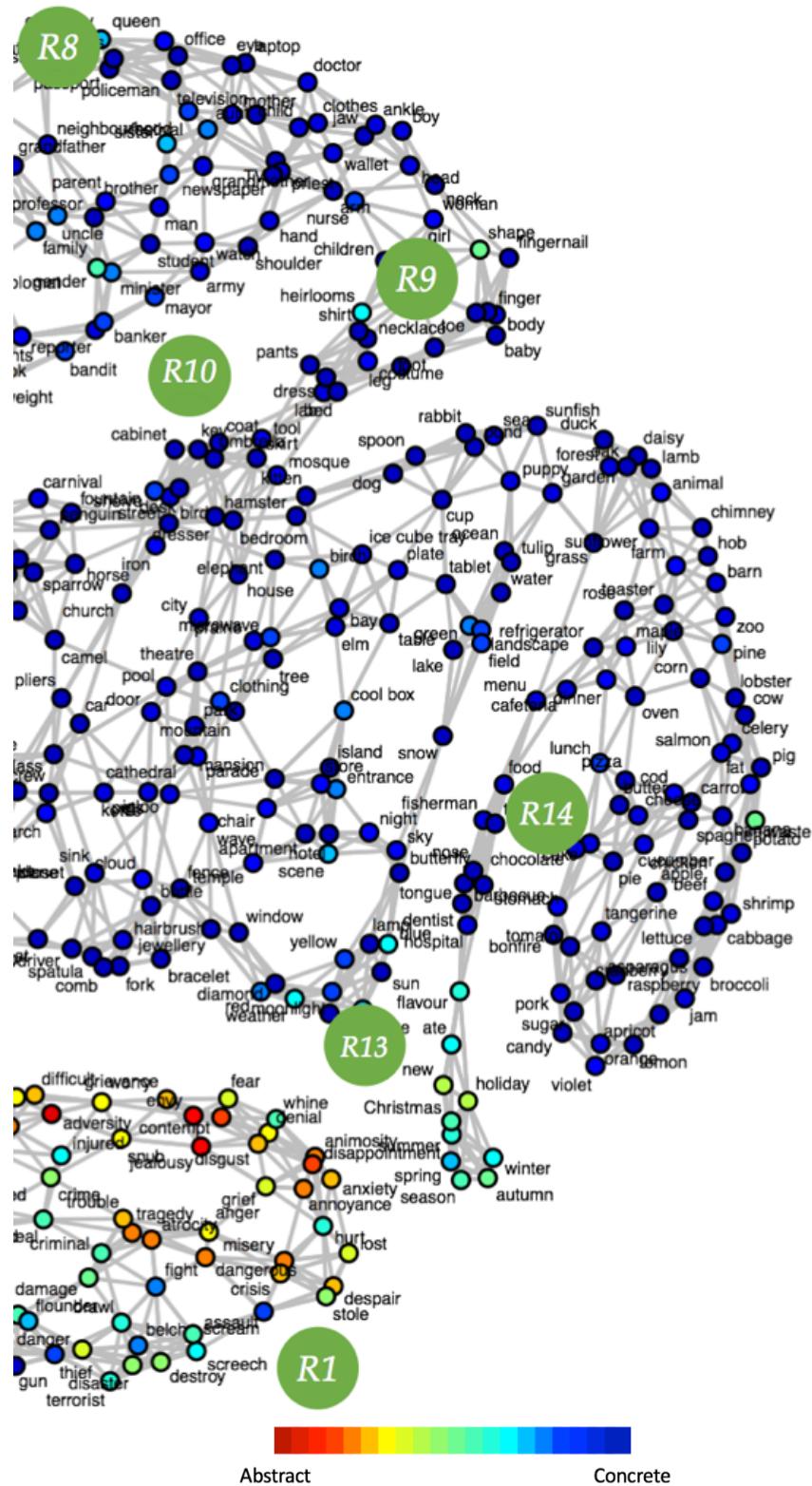


Figure 6.13: Right Panel: Overall semantic topology, including concept labels and the nodes of the network colour-coded according to the concreteness spectrum, ranging from red (abstract) to green (intermediate) and blue (concrete). The green numbered circles (i.e. R1...R14) highlight different portions of the network to aid discussion of more specific network neighbourhoods.

Null Model: Random Topology

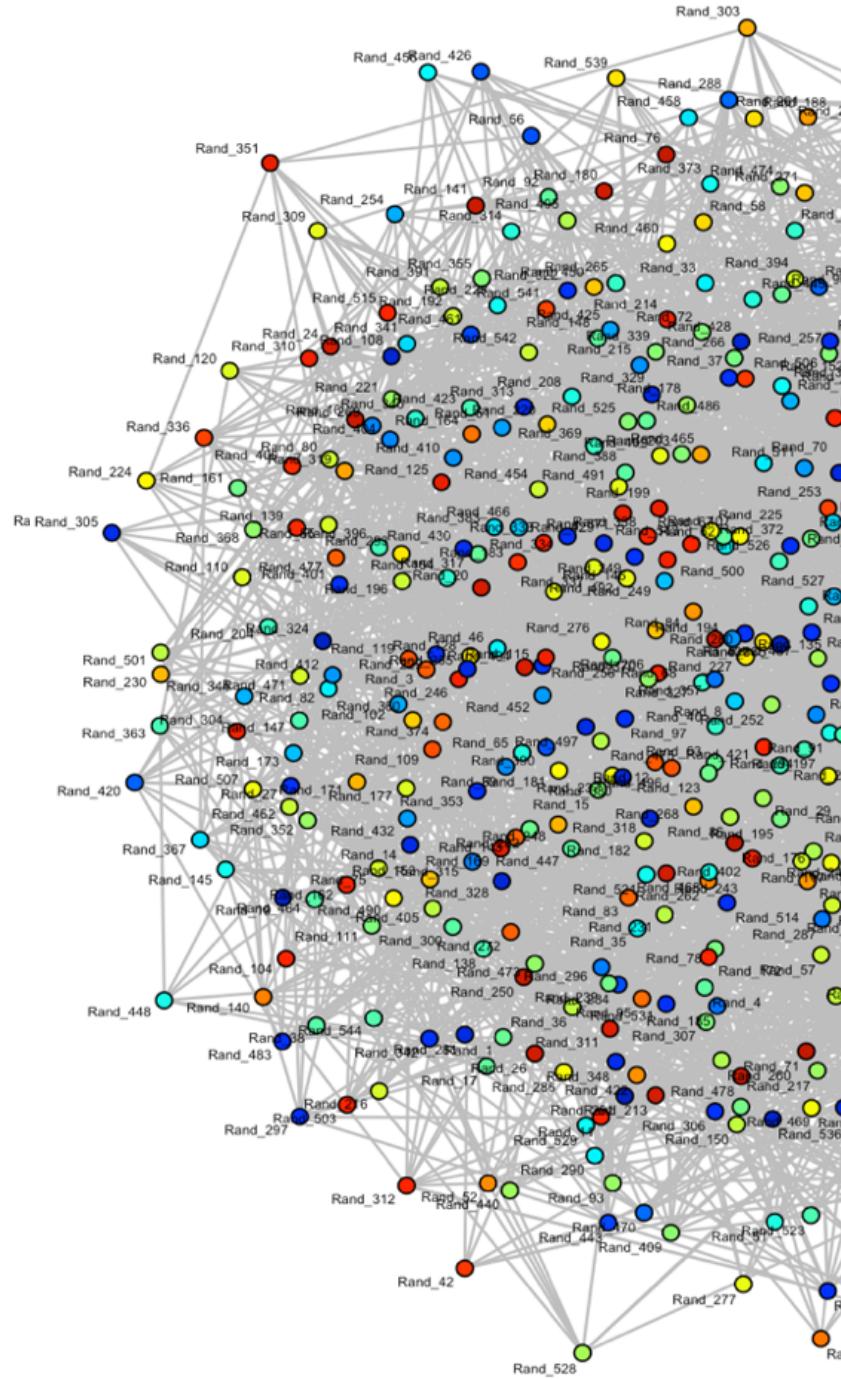


Figure 6.14: Left Panel: A random topology based on shuffling the semantic dimensions across 16 dimensions. This random network topology is generated using the identical set of parameters as our *real* semantic topology. This *null distribution model* also has 544 nodes, 3,264 edges, an *association threshold* ≥ 0.92 , and *t-SNE perplexity* = 60. The node colours are based on the original concreteness spectrum data and as such, is random in this topology.

Null Model: Random Topology

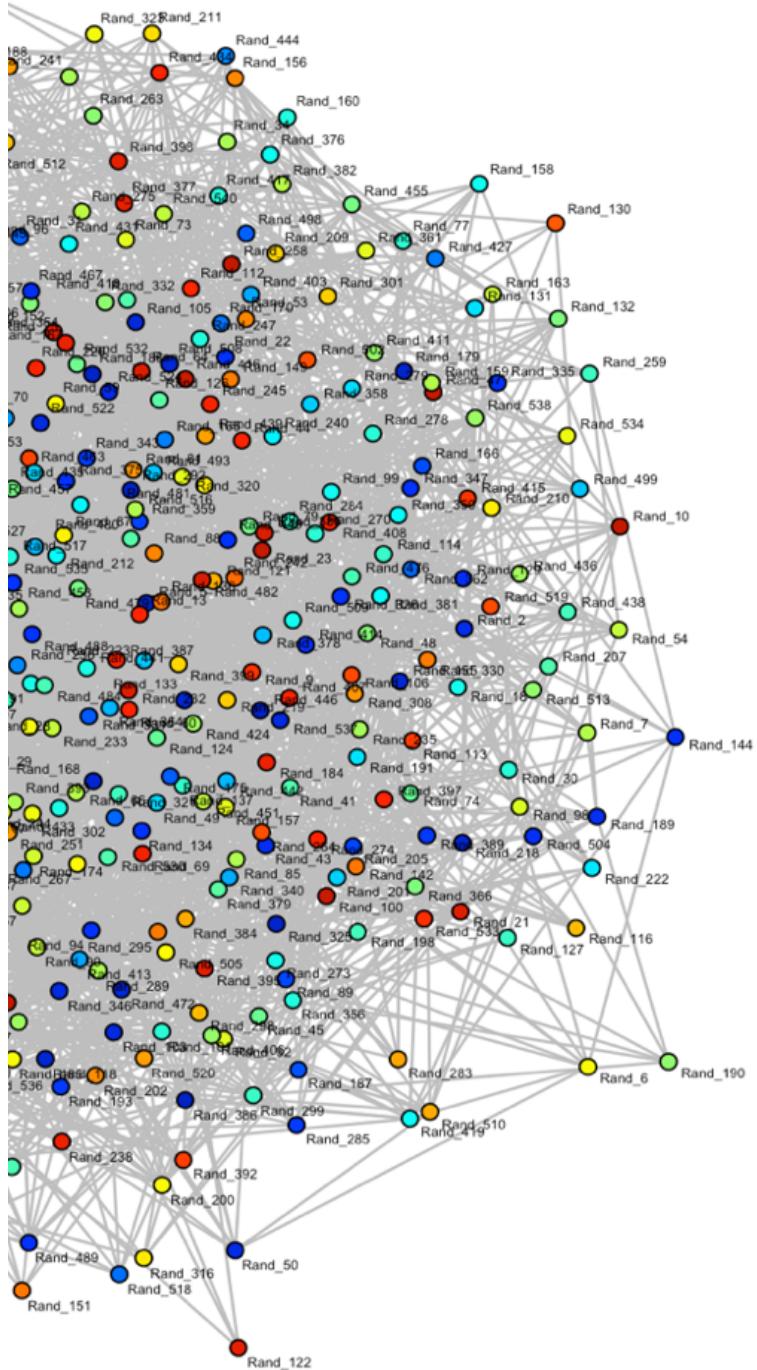
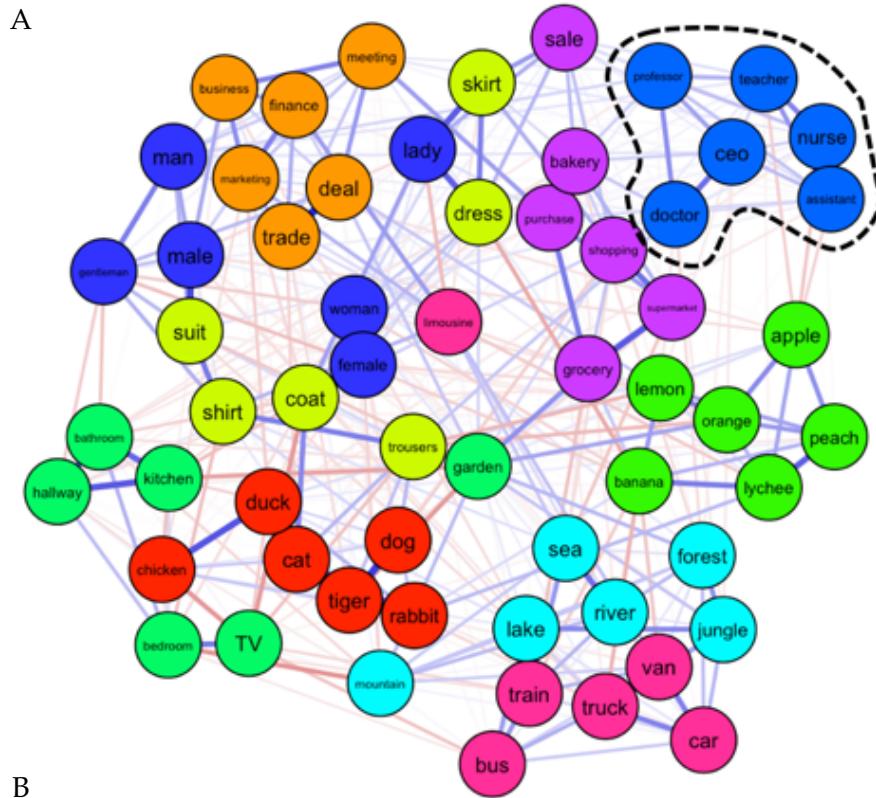


Figure 6.14: Right Panel: A random topology based on shuffling the semantic dimensions across 16 dimensions. This random network topology is generated using the identical set of parameters as our *real* semantic topology. This *null distribution model* also has 544 nodes, 3,264 edges, an *association threshold* ≥ 0.92 , and $t\text{-SNE perplexity} = 60$. The node colours are based on the original concreteness spectrum data and as such, is random in this topology.



B

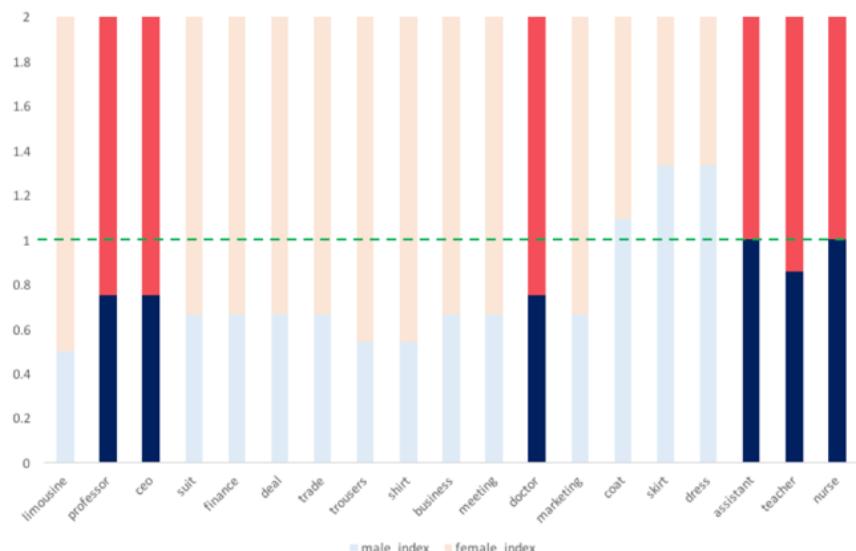
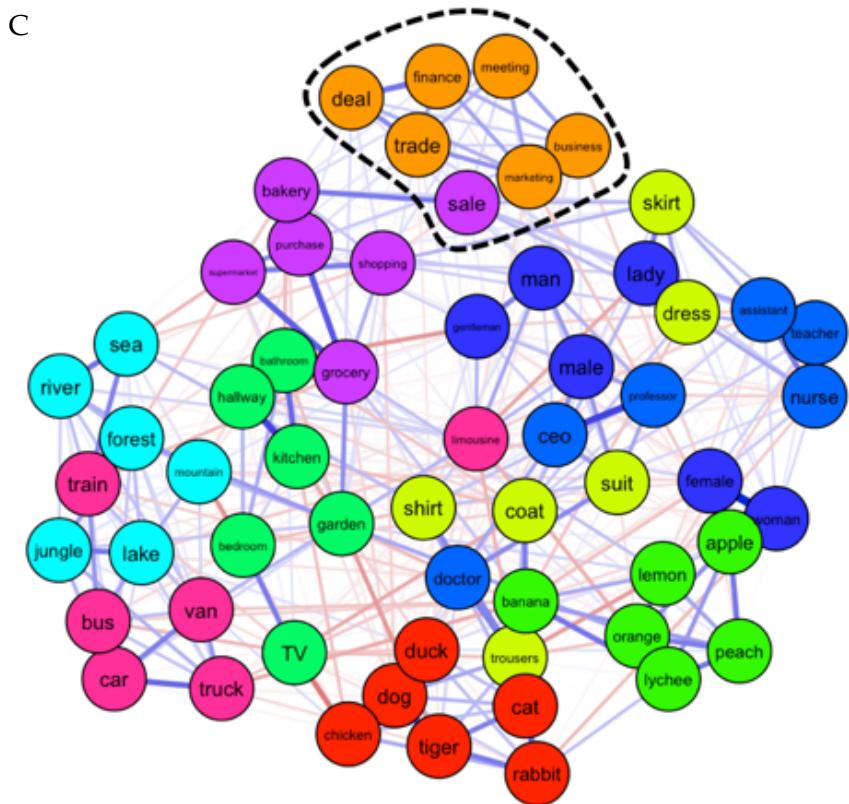


Figure 7.8 A|B: The regularised graphical LASSO network is shown for the condition consisting of debiasing only *occupation* (A). The network has dashed line overlays indicating the specific concepts of interest. The corresponding bias plot is shown in B. The y-axis represents the scaled *gender bias index*. Gender-neutral concepts are indexed at 1 (green dashed line), and more considerable female distances depict bias towards masculine concepts, while larger male distances towards feminine concepts. Darker shades of the colours are used to depict biases of interest for a given experimental condition.



D

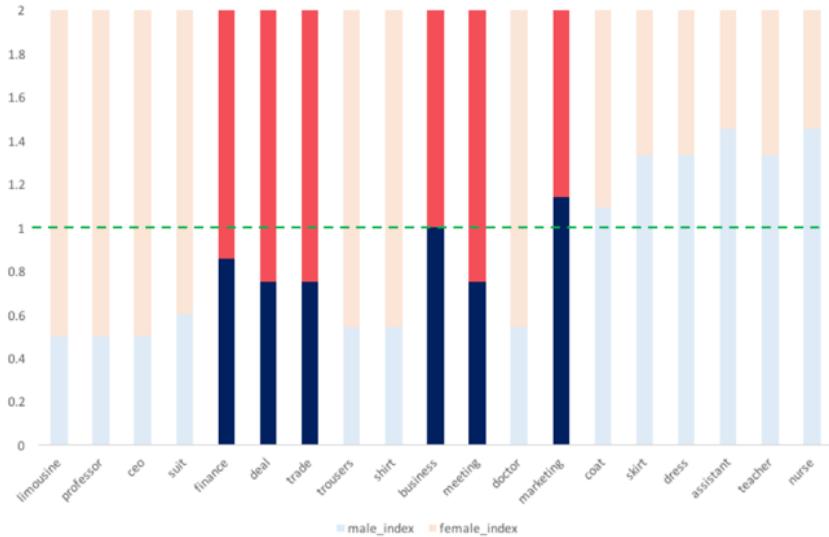
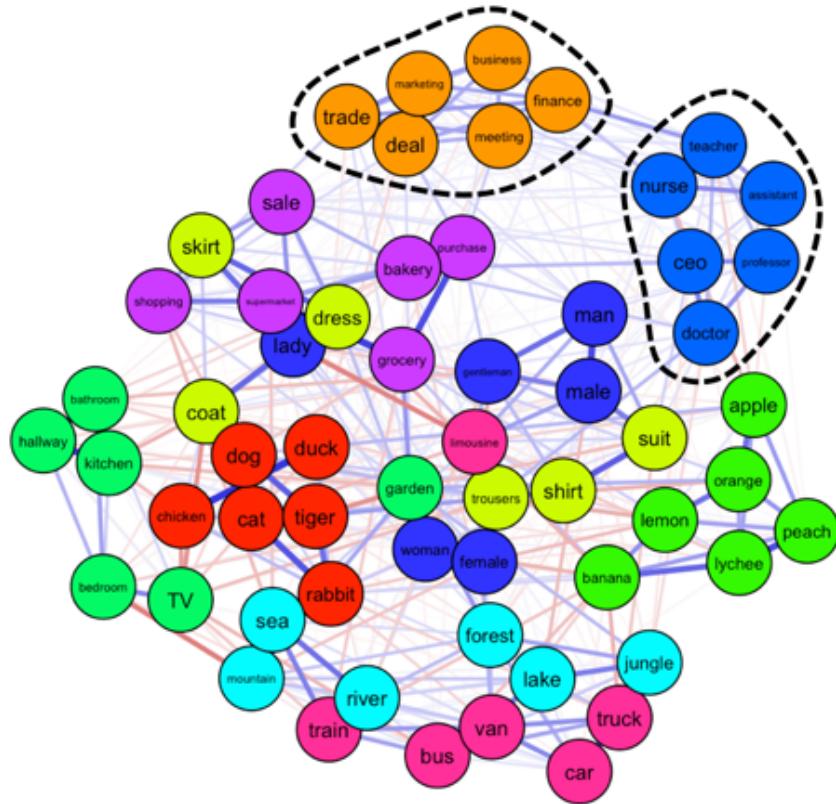


Figure 7.8 C|D: The regularised graphical LASSO network is shown for the condition consisting of debiasing only *business* (C). The network has dashed line overlays indicating the specific concepts of interest. The corresponding bias plot is shown in D. The y-axis represents the scaled *gender bias index*. Gender-neutral concepts are indexed at 1 (green dashed line), and more considerable female distances depict bias towards masculine concepts, while larger male distances towards feminine concepts. Darker shades of the colours are used to depict biases of interest for a given experimental condition.

E



F

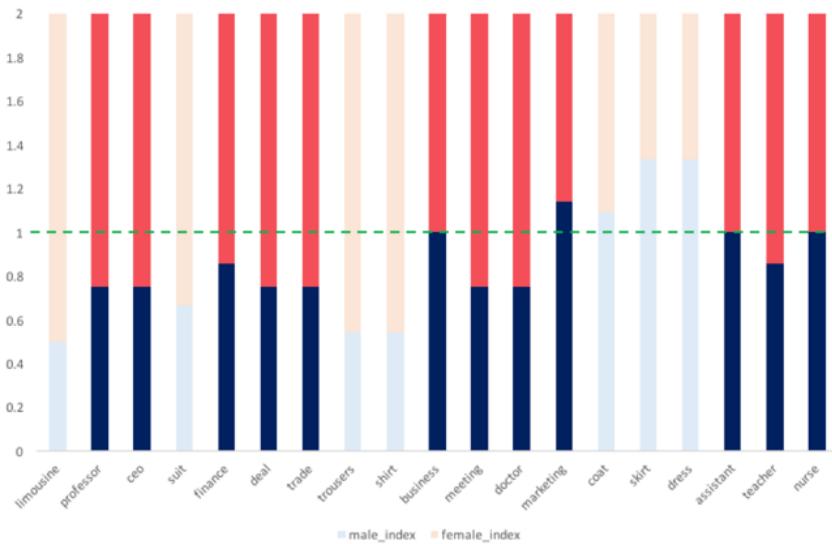


Figure 7.8 E|F: The regularised graphical LASSO network is shown for the condition consisting of debiasing *occupation* and *business* (E). The network has dashed line overlays indicating the specific concepts of interest. The corresponding bias plot is shown in F. The y-axis represents the scaled *gender bias index*. Gender-neutral concepts are indexed at 1 (green dashed line), and more considerable female distances depict bias towards masculine concepts, while larger male distances towards feminine concepts. Darker shades of the colours are used to depict biases of interest for a given experimental condition.

References

- Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9), 814-823.
- Alexander, G. M. (2003). An evolutionary perspective of sex-typed toy preferences: Pink, blue, and the brain. *Archives of Sexual Behavior*, 32(1), 7-14.
- Allen, R., & Hulme, C. (2006). Speech and language processing mechanisms in verbal serial recall. *Journal of Memory and Language*, 55(1), 64-88.
- Altarriba, J., Bauer, L. M., & Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments, & Computers*, 31(4), 578-602.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.
- Anderson, J. R. (2013). *The architecture of cognition*. Hove, UK: Psychology Press.
- Anderson, J. R., & Bower, G.H. (1973). *Human associative memory*. Hove, UK: Psychology Press.
- Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial Intelligence*, 149(1), 91-130.
- Anglin, J. M. (1970). *The Growth of Word Meaning*. Cambridge, MA: MIT Press.
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias risk assessments in criminal sentencing. *ProPublica*. Retrieved from <https://www.propublica.org>
- Asada, M., MacDorman, K. F., Ishiguro, H., & Kuniyoshi, Y. (2001). Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous systems*, 37(2-3), 185-193.
- Ashmore, R. D., & Del Boca, F. K. (1981). Conceptual approaches to stereotypes and stereotyping. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 1-36). Hillsdale, NJ: Erlbaum.
- Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.),

- The psychology of learning and motivation* Vol. 2 (pp. 89–195). Oxford: Academic Press.
- Baddeley, A. (1988). Cognitive psychology and human memory. *Trends in Neurosciences*, 11(4), 176-181.
- Balasubramanian, M., & Schwartz, E. L. (2002). The isomap algorithm and topological stability. *Science*, 295(5552), 7-7.
- Ball, G. H., & Hall, D. J. (1965). *ISODATA: A novel method of data analysis and pattern classification*. Menlo Park, CA: Stanford Research Institute International.
- Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39(3), 445-459.
- Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, 5(8), 617–629.
- Barber, R. F., & Drton, M. (2015). High-dimensional Ising model selection with Bayesian information criteria. *Electronic Journal of Statistics*, 9(1), 567-607.
- Barfuss, W., Massara, G. P., Di Matteo, T., & Aste, T. (2016). Parsimonious modeling with information filtering networks. *Physical Review E*, 94(6), 1-17.
- Barsalou, L. W. (1993). Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A. F. Collins, S. E. Gathercole, M. A. Conway & P. E. Morris (Eds.), *Theories of memory*. (pp. 29-101). Hove, UK: Erlbaum.
- Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and Brain Sciences*, 22(4), 637-660.
- Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, 59, 617-645.
- Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science*, 2(4), 716-724.
- Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thought* (pp. 129–163). New York, NY: Cambridge University Press.
- Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, 7(2), 84-91.
- Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology*, 74(2), 170-179.

- Berger, J., Meredith, M., & Wheeler, S. C. (2008). Contextual priming: Where people vote affects how they vote. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 8846-8849.
- Biederman, I. (1972). Perceiving real-world scenes. *Science*, 177(4043), 77-80.
- Biederman, I. (2017). On the semantics of a glance at a scene. In M. Kubovy, & J. Pomerantz (Eds.), *Perceptual Organization* (pp. 213-253). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2), 143-177.
- Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, 103(3), 597-600.
- Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33(3-4), 130-174.
- Blasch, E., Kadar, I., Salerno, J., Kokar, M. M., Das, S., Powell, G. M., ... & Ruspini, E. H. (2006, May). Issues and challenges of knowledge representation and reasoning methods in situation assessment (Level 2 Fusion). In *Signal Processing, Sensor Fusion, and Target Recognition XV* (Vol. 6235, p. 623510). International Society for Optics and Photonics.
- Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3), 789-865.
- Bloom, L.R. (1998). *Under The Sign of Hope: Feminist Methodology and Narrative Interpretation*. NY: State University of New York Press.
- Bollier, D., & Firestone, C. M. (2010). *The promise and peril of big data* (pp. 1-66). Washington, DC: Aspen Institute, Communications and Society Program.
- Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, 143(3), 263-292.
- Botvinick, M., & Plaut, D. C. (2002). Representing task context: proposals based on a connectionist model of action. *Psychological Research*, 66(4), 298-311.
- Bower, G. H. (1970). Imagery as a relational organizer in associative learning. *Journal of Verbal Learning and Verbal Behavior*, 9(5), 529-533.
- Bowlby, J. (1969). *Attachment* (Attachment and loss series, Vol. 1). New York: Basic Books.

- Brachman, R. J. (1977). What's in a concept: Structural foundations for semantic networks. *International Journal of Man-Machine Studies*, 9(2), 127-152.
- Bream, V., Challacombe, F., Palmer, A., & Salkovskis, P. (2017). *Cognitive Behaviour Therapy for Obsessive-compulsive Disorder*. New York: Oxford University Press.
- Breedin, S. D., Saffran, E. M., & Coslett, H. B. (1994). Reversal of the concreteness effect in a patient with semantic dementia. *Cognitive Neuropsychology*, 11(6), 617-660.
- Broadbent, D. E. (1958). Effect of noise on an "intellectual" task. *The Journal of the Acoustical Society of America*, 30(9), 824-827.
- Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal on Robotics and Automation*, 2(1), 14-23.
- Brooks, R. A. (1989). A robot that walks; emergent behaviors from a carefully evolved network. *Neural Computation*, 1(2), 253-262.
- Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, 6(1-2), 3-15.
- Brooks, R. A. (1991a). Intelligence without reason. *Artificial Intelligence: Critical Concepts*, 3, 107-63.
- Brooks, R. A. (1991b). Intelligence without representation. *Artificial Intelligence*, 47(1-3), 139-159.
- Brooks, R. A. (1991c). New approaches to robotics. *Science*, 253(5025), 1227-1232.
- Brooks, R. A. (1999). *Cambrian intelligence: The early history of the new AI* (Vol. 97). Cambridge, MA: MIT Press.
- Broverman, I. K., Vogel, S. R., Broverman, D. M., Clarkson, F. E., & Rosenkrantz, P. S. (1972). Sex-role stereotypes: A current appraisal1. *Journal of Social Issues*, 28(2), 59-78.
- Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, 84(4), 822-848.
- Brugman, C., & Lakoff, G. (1988). Cognitive topology and lexical networks. In S. Small, G. Cottrell, & M. Tanenhaus (Eds.), *Lexical ambiguity resolution* (pp. 477-507). Palo Alto, CA: Morgan Kaufmann.
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904-911.
- Buckner, R. L., & Krienen, F. M. (2013). The evolution of distributed association networks in the human brain. *Trends in Cognitive Sciences*, 17(12), 648-665.

- Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3), 186-198.
- Bundgaard, P. F., & Østergaard, S. (2007). The story turned upside down: Meaning effects linked to variations on narrative structure. *Semiotica*, 165(1), 263-275.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186.
- Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems*, 30, 3995-4004.
- Cancho, R. F. I., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society B: Biological Sciences*, 268(1482), 2261-2265.
- Cangelosi, A. (2010). Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of Life Reviews*, 7(2), 139-151.
- Cangelosi, A., & Borghi, A. M. (2014). Language and action integration: from humans to cognitive robots. *Topics in Cognitive Science*, 6, 343-558.
- Cangelosi, A., & Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive Science*, 30(4), 673-689.
- Cangelosi, A., & Schlesinger, M. (2015). *Developmental Robotics: From Babies to Robots*. Cambridge, MA: MIT Press.
- Cangelosi, A., & Stramandinoli, F. (2018). A review of abstract concept learning in embodied agents and robots. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1752), 20170131.
- Cangelosi, A., Greco, A., & Harnad, S. (2000). From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. *Connection Science*, 12(2), 143-162.
- Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience*, 10(1), 1-34.
- Chemero, A. (2011). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.
- Cheryan, S., Master, A., & Meltzoff, A. N. (2015). Cultural stereotypes as gatekeepers: Increasing girls' interest in computer science and engineering by diversifying stereotypes. *Frontiers in Psychology*, 6(49), 1-8.

- Chodorow, N. J. (1995). Gender as a personal and cultural construction. *Signs: Journal of Women in Culture and Society*, 20(3), 516-544.
- Chomsky, N. (1957). Logical structures in language. *American Documentation*, 8(4), 284-291.
- Christensen, A. P., Kenett, Y. N., Aste, T., Silvia, P. J., & Kwapił, T. R. (2018). Network structure of the Wisconsin Schizotypy Scales-Short Forms: Examining psychometric network filtering approaches. *Behavior Research Methods*, 50(6), 2531-2550.
- Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5), 489-509.
- Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, 4(5), 170-178.
- Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, 36(1), 28-71.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284-290.
- Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York: Oxford University Press.
- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6), 407-428.
- Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2), 240-247.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12, 2493-2537.
- Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4), 497-505.
- Connell, L., Lynott, D., & Carney, J. (2017). Interoception: The Forgotten Modality in Perceptual Grounding of Concepts. In *Proceedings of Cognitive Science Society*.
- Cooper, R., & Shallice, T. (2006). Hierarchical schemas and goals in the control of sequential behavior. *Psychological Review*, 113(4), 887-916.
- Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, 17(4), 297-338.
- Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Möttus, R., Waldorp, L. J., & Cramer, A. O. (2015). State of the art personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, 54, 13-29.

- Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10(7), 1-9.
- Cowell, F. A. (2000). Measurement of inequality. In A. Atkinson & F. Bourguignon (Eds.), *Handbook of Income Distribution* (pp. 87-166). Amsterdam: Elsevier Science.
- Cramer, P. (1970). A study of homographs. In L. Postman & G. Keppel (Eds.), *Norms of Word Association* (pp. 361-382). New York: Academic Press.
- Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132(2), 163-201.
- Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6), 453-482.
- Crutch, S. J., Troche, J., Reilly, J., & Ridgway, G. R. (2013). Abstract conceptual feature ratings: the role of emotion, magnitude, and other cognitive domains in the organization of abstract conceptual knowledge. *Frontiers in Human Neuroscience*, 7, 186, 1-14.
- Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., & Damasio, A. R. (1996). A neural basis for lexical retrieval. *Nature*, 380(6574), 499-505.
- Damasio, H., Tranel, D., Grabowski, T., Adolphs, R., & Damasio, A. (2004). Neural systems behind word and concept retrieval. *Cognition*, 92(1-2), 179-229.
- Dancy, C. L. (2013). ACT-R Φ : A cognitive architecture with physiology and affect. *Biologically Inspired Cognitive Architectures*, 6, 40-45.
- De La Cruz, V. M., Di Nuovo, A., Di Nuovo, S., & Cangelosi, A. (2014). Making fingers and words count in a cognitive robot. *Frontiers in Behavioral Neuroscience*, 8(13), 1-12.
- Dennett, D. C. (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness*. Cambridge, MA: MIT Press.
- Dennett, D. C. (2006). The frame problem of AI. *Philosophy of Psychology: Contemporary Readings*, 433, 67-83.
- Dennis, J. M. (2001). Are Internet panels creating professional respondents?. *Marketing Research*, 13(2), 34-38.
- Dolan, R. J., & Vuilleumier, P. (2003). Amygdala automaticity in emotional processing. *Annals of the New York Academy of Sciences*, 985(1), 348-355.
- Dorfman, P. W. (1979). Measurement and meaning of recreation satisfaction: A case study in camping. *Environment and Behavior*, 11(4), 483-510.
- Dove, G. (2011). On the need for embodied and dis-embodied cognition. *Frontiers in Psychology*, 1(242), 1-13.

- Dreyfus, H. L. (1997). Heidegger on Gaining a Free Relation to Technology. In K. Schrader-Frechette & L. Westra (Eds.), *Technology and Values* (pp. 41-53). Lanham, MD: Rowman & Littlefield.
- Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philosophical Psychology*, 20(2), 247-268.
- Ebeling, C. L. (1978). *Syntax and Semantics: A taxonomic approach*. Leiden: Brill.
- Egstrom, G. H., Weltman, G., Baddeley, A. D., Cuccaro, W. J., & Willis, M. A. (1972). *Underwater work performance and work tolerance*. Los Angeles, Calif: University of California, School of Engineering and Applied Sciences; 1972. UCLA-ENG-7243. Biotechnological Laboratory Technical Report 51.
- Eide, E., & Gish, H. (1996). A parametric approach to vocal tract length normalization. In *Proceedings of ICASSP*.
- Ellis, R., & Humphreys, G. W. (1999). *Connectionist psychology: A text with readings*. Hove, UK: Psychology Press.
- Epskamp, S., & Fried, E. I. (2016). *A tutorial on regularized partial correlation networks*. Retrieved from <https://arxiv.org/abs/1607.01367>.
- Etkin, A., Egner, T., Peraza, D. M., Kandel, E. R., & Hirsch, J. (2006). Resolving emotional conflict: a role for the rostral anterior cingulate cortex in modulating activity in the amygdala. *Neuron*, 51(6), 871-882.
- Euler, L. (1736). *Mechanica sive motus scientia analytice exposita*. St Petersburg, Russia: Ex typographia Academiae scientiarum.
- Everett, D., Berlin, B., Gonçalves, M., Kay, P., Levinson, S., Pawley, A., ... & Everett, D. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology*, 46(4), 621-646.
- Eysenck, M. W., & Keane, M. T. (2005). *Cognitive psychology: A student's handbook*. Hove, UK: Psychology Press.
- Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, 120(4), 339-357.
- Farkaš, I., Malík, T., & Rebrová, K. (2012). Grounding the meanings in sensorimotor behavior using reinforcement learning. *Frontiers in Neurorobotics*, 6, 1-13.
- Fausto-Sterling, A. (2012). *Sex/gender: Biology in a social world*. London: Routledge.
- Fay, A. J., & Maner, J. K. (2012). Warmth, spatial proximity, and social attachment: The embodied perception of a social metaphor. *Journal of Experimental Social Psychology*, 48(6), 1369-1372.

- Feldman, J., & Narayanan, S. (2004). Embodied meaning in a neural theory of language. *Brain and Language*, 89(2), 385-392.
- Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. New York: Oxford University Press.
- Fodor, J. A., & Pylyshyn, Z. W. (1981). How direct is visual perception? Some reflections on Gibson's "ecological approach.". *Cognition*, 9(2), 139-196.
- Fodor, J. D. (1983). Phrase structure parsing and the island constraints. *Linguistics and Philosophy*, 6(2), 163-223.
- Forbes, M. K., Wright, A. G., Markon, K. E., & Krueger, R. F. (2017). Evidence that psychopathology symptom networks have limited replicability. *Journal of Abnormal Psychology*, 126(7), 969-988.
- Foygel, R., & Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems* (pp. 604-612).
- Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. *Cognition*, 108(3), 819-824.
- Franklin, S. (2014). History, motivations, and core themes. In K. Frankish & W. Ramsey (Eds.). *The Cambridge Handbook of Artificial Intelligence* (pp. 15-33). New York, NY: Cambridge University Press.
- Fraser, H., & Stevenson, B. (2014). The power and persistence of contextual priming: more risks in using police transcripts to aid jurors' perception of poor quality covert recordings. *The International Journal of Evidence & Proof*, 18(3), 205-229.
- Froese, T. (2007). On the role of AI in the ongoing paradigm shift within the cognitive sciences. In M. Lungarella, et al. (Eds.), *50 Years of Artificial Intelligence* (pp. 63-75). Berlin: Springer-Verlag.
- Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, 173(3-4), 466-500.
- Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11), 1129-1164.
- Gabbe, M. (2016). Aristotle on the Metaphysics of Emotions. *Apeiron*, 49(1), 33-56.
- Gallese, V., & Cuccio, V. (2018). The neural exploitation hypothesis and its implications for an embodied approach to language and cognition: Insights from the study of action verbs processing and motor disorders in Parkinson's disease. *Cortex*, 100, 215-225.

- Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, 22(3-4), 455-479.
- Garbarini, F., & Adenzato, M. (2004). At the root of embodied cognition: Cognitive science meets neurophysiology. *Brain and Cognition*, 56(1), 100-106.
- Garzón, P., & Keijzer, F. (2011). Plants: Adaptive behavior, root-brains, and minimal cognition. *Adaptive Behavior*, 19(3), 155-171.
- Georgeon, O. L., & Cordier, A. (2014). Inverting the interaction cycle to model embodied agents. *Procedia Computer Science*, 41, 243-248.
- Gibbs, J. L., Ellison, N. B., & Heino, R. D. (2006). Self-presentation in online personals: The role of anticipated future interaction, self-disclosure, and perceived success in Internet dating. *Communication Research*, 33(2), 152-177.
- Gibney, E. (2015). DeepMind algorithm beats people at classic video games. *Nature*, 518(7540), 465-466.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.
- Gigerenzer, G. (2009). Surrogates for theory. *Observer*, 22(2), 21-23.
- Gilbert, A. L., Regier, T., Kay, P., & Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences*, 103(2), 489-494.
- Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, 12(4), 395-427.
- Gini, C. (1912). Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T)*. Rome: Libreria Eredi Virgilio Veschi.
- Gipper, H. (1972). Gibt es ein Sprachliches Relativitätsprinzip? Untersuchungen zur Sapir-Whorf-Hypothese. Frankfurt am Main, Germany: S. Fischer.
- Gisbrecht, A., Schulz, A., & Hammer, B. (2015). Parametric nonlinear dimensionality reduction using kernel t-SNE. *Neurocomputing*, 147, 71-82.
- Glenberg, A. M. (2015). Few believe the world is flat: How embodiment is changing the scientific understanding of cognition. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 69(2), 165-171.
- Glenberg, A. M., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., & Buccino, G. (2008). Processing abstract language modulates motor system activity. *The Quarterly Journal of Experimental Psychology*, 61(6), 905-919.

- Goldstone, R. L., & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, 65(2-3), 231-262.
- Goldstone, R. L., & Rogosky, B. J. (2002). Using relations within conceptual systems to translate across conceptual systems. *Cognition*, 84(3), 295-320.
- Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). Cambridge: MIT Press.
- Gordon, B. (1985). Subjective frequency and the lexical decision latency function: Implications for mechanisms of lexical access. *Journal of Memory and Language*, 24(6), 631-645.
- Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, 306(5695), 496-499.
- Gottesman, C. V., & Intraub, H. (1999). Wide-angle memories of close-up scenes: A demonstration of boundary extension. *Behavior Research Methods, Instruments, & Computers*, 31(1), 86-93.
- Grabowski, T. J., Damasio, H., & Damasio, A. R. (1998). Premotor and prefrontal correlates of category-related lexical retrieval. *Neuroimage*, 7(3), 232-243.
- Graffigna, G., Barella, S., Bonanomi, A., & Lozza, E. (2015). Measuring patient engagement: development and psychometric properties of the Patient Health Engagement (PHE) scale. *Frontiers in Psychology*, 6, 274, 1-10.
- Greco, A., Riga, T., & Cangelosi, A. (2003). The acquisition of new categories through grounded symbols: An extended connectionist model. In *Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003* (pp. 763-770). Berlin: Springer-Verlag.
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1), 4-27.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, 74(6), 1464-1480.
- Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85(2), 197-216.
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97(1), 17-41.
- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114(2), 211-244.

- Grondin, R., Lupker, S. J., & McRae, K. (2009). Shared features dominate semantic richness effects for concrete concepts. *Journal of Memory and Language*, 60(1), 1-19.
- Gunning, D. (2016). Explainable Artificial Intelligence (XAI): Technical report Defense Advanced Research Projects Agency DARPA-BAA-16-53. *DARPA, Arlington, USA*.
- Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *The Quarterly Journal of Experimental Psychology*, 69(4), 626-653.
- Hadid, A., Kouropeteva, O., & Pietikainen, M. (2002). Unsupervised learning using locally linear embedding: experiments with face pose analysis. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on* (Vol. 1, pp. 111-114). IEEE.
- Hahn, U. (2011). The problem of circularity in evidence, argument, and explanation. *Perspectives on Psychological Science*, 6(2), 172-182.
- Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). Cluster validity methods: part I. *ACM Sigmod Record*, 31(2), 40-45.
- Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4), 835-845.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (pp. 3315-3323).
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3), 335-346.
- Harnad, S. (1992). Connecting object to symbol in modelling cognition. In *Connectionism in Context* (pp. 75-90). London: Springer.
- Harquail, C. V., & King, A. W. (2010). Construing organizational identity: The role of embodied cognition. *Organization Studies*, 31(12), 1619-1648.
- Haugeland, J. (1985). *Artificial Intelligence: The very idea*. Cambridge, MA: MIT Press.
- Haugeland, J. (1993). *Mind embodied and embedded*. Cambridge, MA: Harvard University Press.
- Hauk, O. (2004). Keep it simple: a case for using classical minimum norm estimation in the analysis of EEG and MEG data. *Neuroimage*, 21(4), 1612-1621.
- Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. London: Chapman & Hall.
- Heidorn, G. E. (1975). Augmented phrase structure grammars. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing* (pp. 1-5). Association for Computational Linguistics.

- Hellström, N. P. (2012). Darwin and the Tree of Life: The roots of the evolutionary tree. *Archives of Natural History*, 39(2), 234-252.
- Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, 98(1), 74-95.
- Hinton, G. H. (1981). Implementing semantic networks in parallel hardware. In G. H. Hinton & J. A. Anderson (Eds.), *Parallel Models of Associative Memory*. Hillsdale, NJ: Erlbaum.
- Hoffman, P., McClelland, J. L., Ralph, L., & Matthew, A. (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological Review*, 125(3), 293-328.
- Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, 45(3), 718-730.
- Hölldobler, S., Kalinke, Y., & Störr, H. P. (1999). Approximating the semantics of logic programs by recurrent neural networks. *Applied Intelligence*, 11(1), 45-58.
- Holmes, K. J., & Wolff, P. (2010, January). Simulation from schematics: Dorsal stream processing and the perception of implied motion. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32, No. 32).
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193-218.
- Humphries, M. D., & Gurney, K. (2008). Network ‘small-world-ness’: a quantitative method for determining canonical network equivalence. *PloS One*, 3(4), 1-10.
- Hunt, T., Song, C., Shokri, R., Shmatikov, V., & Witchel, E. (2018). Chiron: Privacy-preserving Machine Learning as a Service. *ArXiv Preprint ArXiv:1803.05961*.
- Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600), 453-458.
- Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6), 1210-1224.
- i Cancho, R. F., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1482), 2261-2265.
- Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology*, 60, 653-670.

- Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. *Cognitive Development*, 9(1), 45-75.
- Inglehart, R., & Norris, P. (2016). Trump, Brexit, and the rise of populism: Economic have-nots and cultural backlash. In *Presidential plenary panel during 24th World Congress of the International Political Science Association*, Poznan, 25 July.
- Ingram, S., & Munzner, T. (2015). Dimensionality reduction for documents with nearest neighbor queries. *Neurocomputing*, 150, 557-569.
- Intraub, H., Gottesman, C. V., Willey, E. V., & Zuk, I. J. (1996). Boundary extension for briefly glimpsed photographs: Do common perceptual processes result in unexpected memory distortions?. *Journal of Memory and Language*, 35(2), 118-134.
- Jackendoff, R. (2002). The mind doesn't work that way: the scope and limits of computational psychology. *Language*, 78(1), 164-170.
- Jacobs, R. A., & Kosslyn, S. M. (1994). Encoding shape and spatial relations: The role of receptive field size in coordinating complementary representations. *Cognitive Science*, 18(3), 361-386.
- Jarratt, S. C. (1998). *Rereading the sophists: Classical rhetoric refigured*. Springfield, IL: SIU Press.
- Jaworska-Biskup, K. (2011). The world without sight. A comparative study of concept understanding in Polish congenitally totally blind and sighted children. *Psychology of Language and Communication*, 15(1), 27-48.
- Jenset, G. B., & McGillivray, B. (2017). *Quantitative Historical Linguistics: A Corpus Framework* (Vol. 26). New York: Oxford University Press.
- Jim, K. C., Giles, C. L., & Horne, B. G. (1996). An analysis of noise in recurrent neural networks: convergence and generalization. *IEEE Transactions on Neural Networks*, 7(6), 1424-1438.
- Johnson-Laird, P. N., & Shafir, E. (1993). The interaction between reasoning and decision making: An introduction. *Cognition*, 49(1-2), 1-9.
- Johnson-Laird, P. N., Herrmann, D. J., & Chaffin, R. (1984). Only connections: A critique of semantic networks. *Psychological Bulletin*, 96(2), 292-315.
- Johnson, M. (2013). *The body in the mind: The bodily basis of meaning, imagination, and reason*. Chicago, IL: University of Chicago Press.
- Johnson, M. (1981). *Philosophical Perspectives on Metaphor*. Minneapolis, MN: University of Minnesota Press.
- Johnson, M. G., & Henley, T. B. (1988). Something Old, Something New, Something Borrowed, Something True. *Metaphor and Symbol*, 3(4), 233-252.

- Johnson, M. G., & Malgady, R. G. (1979). Some cognitive aspects of figurative language: Association and metaphor. *Journal of Psycholinguistic Research*, 8(3), 249-265.
- Jolliffe, I. (2011). Principal component analysis. In *International encyclopedia of statistical science* (pp. 1094-1096). Berlin: Springer-Verlag.
- Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, 55(4), 534-552.
- Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2016). Rawlsian fairness for machine learning. *ArXiv Preprint ArXiv:1610.09559*.
- Kahneman, D. (2011). Don't Blink! The Hazards of Confidence. *New York Times*, 19.
- Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3), 187-200.
- Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1), 7-15.
- Kandola, B., & Kandola, J. (2013). *The Invention of Difference: The Story of Gender Bias at Work*. London, UK: BookBaby.
- Kiela, D. (2017). *Deep embodiment: grounding semantics in perceptual modalities* (No. UCAM-CL-TR-899). University of Cambridge, Computer Laboratory.
- Kiela, D., & Clark, S. (2017). Learning neural audio embeddings for grounding semantics in auditory perception. *Journal of Artificial Intelligence Research*, 60, 1003-1030.
- Kiela, D., Bulat, L., Vero, A. L., & Clark, S. (2016). Virtual embodiment: A scalable long-term strategy for artificial intelligence research. *ArXiv Preprint ArXiv:1610.07432*.
- Kieras, D. (1978). Beyond pictures and words: Alternative information-processing models for imagery effect in verbal memory. *Psychological Bulletin*, 85(3), 532-554.
- Kilbourne, B. S., England, P., Farkas, G., Beron, K., & Weir, D. (1994). Returns to skill, compensating differentials, and gender bias: Effects of occupational characteristics on the wages of white women and men. *American Journal of Sociology*, 100(3), 689-719.
- Kirsh, D. (1991). Foundations of AI: the big issues. *Artificial Intelligence*, 47(1-3), 3-30.
- Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9), 1464-1480.
- Kolosnjaji, B., Zarras, A., Webster, G., & Eckert, C. (2016, December). Deep learning for classification of malware system call sequences.

- In *Australasian Joint Conference on Artificial Intelligence* (pp. 137-149). Springer, Cham.
- Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: why emotion matters. *Journal of Experimental Psychology: General*, 140(1), 14-34.
- Kousta, S. T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, 112(3), 473-481.
- Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6), 1126-1141.
- Laird, J. E. (2008). Extending the Soar cognitive architecture. *Frontiers in Artificial Intelligence and Applications*, 171, 224-235.
- Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and thought* (pp. 202-251). New York, NY: Cambridge University Press.
- Lakoff, G. (2008). *Women, fire, and dangerous things*. Chicago, IL: University of Chicago Press.
- Lakoff, G. (2016). Language and emotion. *Emotion Review*, 8(3), 269-273.
- Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive Science*, 4(2), 195-208.
- Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh* (Vol. 4). New York: Basic Books.
- Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. Chicago, IL: University of Chicago Press.
- Lallée, S., Lemaignan, S., Lenz, A., Melhuish, C., Natale, L., Skachek, S., ... & Dominey, P. F. (2010, October). Towards a platform-independent cooperative human-robot interaction system: I. perception. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on* (pp. 4444-4451). IEEE.
- Landau, B., Smith, L. B., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3(3), 299-321.
- Landau, B., Smith, L. B., & Jones, S. (1992). Syntactic context and the shape bias in children's and adults' lexical learning. *Journal of Memory and Language*, 31(6), 807-825.
- Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. *Psychology of Learning and Motivation*, 41, 43-84.

- 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., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3), 259-284.
- Landauer, T., McNamara, D., Dennis, S., & Kintsch, W. (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.
- Laurence, S., & Margolis, E. (1999). *Concepts and Cognitive Science*. Cambridge, MA: MIT Press.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
- Lenat, D. B., Feigenbaum, E. A., Smith, B. C., Rosenbloom, P. S., Laird, J. E., Newell, A., ... & Norman, D. A. (1991). 10. On the thresholds of knowledge. *Artificial Intelligence*, 47(1-3), 185-250.
- Lipovetsky, S., & Conklin, W. M. (2015). Predictor relative importance and matching regression parameters. *Journal of Applied Statistics*, 42(5), 1017-1031.
- Lippmann, W. (1922). Stereotypes. *Public Opinion and the Press*. New York: Macmillan Publishers.
- Liu, H., & Singh, P. (2004). ConceptNet—A practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4), 211-226.
- Louviere, J. J. (1992). Experimental choice analysis: Introduction and overview. *Journal of Business Research*, 24(2), 89-95.
- Louviere, J. J., Finn, A., & Timmermans, H. G. (1994). *Retail Research Methods, Handbook of Marketing Research*. New York: McGraw-Hill.
- Louviere, J., Swait, J., & Andreson, D. (1995). *Best-Worst Conjoint: A New Preference Elicitation Method to Simultaneously Identify Overall Importance and Attribute Level Partworth*. Working Paper. University of Florida, Gainesville, FL.
- Louwerse, M. M. (2007). Symbolic or embodied representations: A case for symbol interdependency. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 107–120). Mahwah, NJ: Erlbaum.
- Louwerse, M. M. (2008). Embodied relations are encoded in language. *Psychonomic Bulletin and Review*, 15(4), 838-844.
- Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2), 273-302.
- Louwerse, M. M., & Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. *Cognition*, 114(1), 96-104.
- Ludwig, K., & Schneider, S. (2008). Fodor's challenge to the classical computational theory of mind. *Mind & Language*, 23(1), 123-143.

- Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2), 203-208.
- Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior Research Methods*, 45(2), 516-526.
- Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris*, 102(1-3), 59-70.
- Mainzer, K. (2009). From embodied mind to embodied robotics: Humanities and system theoretical aspects. *Journal of Physiology-Paris*, 103(3-5), 296-304.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. New York, NY: Cambridge University Press.
- Markman, A. B., & Brendl, C. M. (2005). Constraining theories of embodied cognition. *Psychological Science*, 16(1), 6-10.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: WH Freeman.
- Mataric, M. J. (1990). A distributed model for mobile robot environment-learning and navigation. *Technical Report AI-TR-1228*, MIT.
- Mataric, M. J., & Brooks, R. A. (1990, July). Learning a distributed map representation based on navigation behaviors. In *Proceedings of 1990 USA Japan Symposium on Flexible Automation* (pp. 499-506).
- Matheson, H. E., & Barsalou, L. W. (2017) Embodied cognition. In: J. Wixted (Ed.) *The Stevens' handbook of experimental psychology and cognitive neuroscience* (4th ed.). Hoboken: Wiley.
- McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4, 463-502.
- McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, 4(4), 310-322.
- McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1988). The appeal of parallel distributed processing. In A. M. Collins & E. E. Smith (Eds.), *Readings in cognitive science: A perspective from psychology and artificial intelligence* (pp. 52-72). San Mateo, CA: Morgan Kaufmann.
- McFadden, D. (1977). *Quantitative methods for analyzing travel behavior of individuals: Some recent developments*. Institute of Transportation Studies, University of California.
- McNally, R. J., Robinaugh, D. J., Wu, G. W., Wang, L., Deserno, M. K., & Borsboom, D. (2015). Mental disorders as causal systems: A network

- approach to posttraumatic stress disorder. *Clinical Psychological Science*, 3(6), 836-849.
- McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. Hove, UK: Psychology Press.
- McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4), 547-559.
- Mercier, E. M., Barron, B., & O'connor, K. M. (2006). Images of self and others as computer users: The role of gender and experience. *Journal of Computer Assisted Learning*, 22(5), 335-348.
- Merleau-Ponty, M. (1945). *Phenomenology of Perception*. Paris, France: Gallimard.
- Merrell, F. (1997). *Peirce, signs, and meaning*. Toronto, Canada: University of Toronto Press.
- Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, 48(7), 788-804.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111-3119).
- Minato, T., & Ishiguro, H. (2007, October). Generating natural posture in an android by mapping human posture in three-dimensional position space. In *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on* (pp. 609-616). IEEE.
- Mirolli, M., & Parisi, D. (2009). Language as a cognitive tool. *Minds and Machines*, 19(4), 517-528.
- Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880), 1191-1195.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533.
- Moffitt, B. (2016). *The global rise of populism: Performance, political style, and representation*. Stanford, US: Stanford University Press.
- Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, 96(2), 143-182.
- Morton, J. B., & Trehub, S. E. (2001). Children's understanding of emotion in speech. *Child Development*, 72(3), 834-843.

- Munson, J. H. (1971). *Robot planning, execution, and monitoring in an uncertain environment* (No. SRI-TN-59). Stanford Research Institute, Menlo Park, CA.
- Mwangi, B., Soares, J. C., & Hasan, K. M. (2014). Visualization and unsupervised predictive clustering of high-dimensional multimodal neuroimaging data. *Journal of Neuroscience Methods*, 236, 19-25.
- Myers, D. G. (1995). *Psychology. Fourth Edition*. New York, NY: Worth.
- Nagy, G., Seth, S. C., & Stoddard, S. D. (1986). Document analysis with an expert system. In *Proceedings of Pattern Recognition in Practice, Volume II* (pp. 149-159). Amsterdam.
- Narayanan, S. (1997, August). Talking the talk is like walking the walk: A computational model of verbal aspect. In *Proceedings of the 19th Cognitive Science Society Conference* (pp. 548-553). Hillsdale, NJ: Erlbaum.
- Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, 106(3), 226-254.
- Newell, A. (1980). Physical Symbol Systems. *Cognitive Science*, 4(2), 135-183.
- Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-Hall.
- Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry. *Communications of the ACM*, 19, 113-126.
- Nicodemus, K. K., Elvevåg, B., Foltz, P. W., Rosenstein, M., Diaz-Asper, C., & Weinberger, D. R. (2014). Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach. *Cortex*, 55, 182-191.
- Niedenthal, P. M. (2007). Embodying Emotion. *Science*, 316(5827), 1002-1005.
- Niedenthal, P. M., Barsalou, L. W., Winkielman, P., Krauth-Gruber, S., & Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and Social Psychology Review*, 9(3), 184-211.
- Niedenthal, P. M., Brauer, M., Halberstadt, J. B., & Innes-Ker, Å. H. (2001). When did her smile drop? Facial mimicry and the influences of emotional state on the detection of change in emotional expression. *Cognition & Emotion*, 15(6), 853-864.
- Nolfi, S., Bongard, J., Husbands, P., & Floreano, D. (2016). *Evolutionary Robotics*. Cambridge, MA: MIT Press.
- Norvig, P. (1992). *Paradigms of artificial intelligence programming: Case studies in Common LISP*. San Francisco, CA: Morgan Kaufmann.
- Nosek, B. A. (2007). Implicit-explicit relations. *Current Directions in Psychological Science*, 16(2), 65-69.

- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1), 101-115.
- Ogden, C. K., & Richards, I. A. (1923). *The Meaning of Meaning*. New York, NY: Harcourt, Brace & World, Inc.
- Orme, B. K. (2009). Anchored Scaling in MaxDiff Using Dual Response. *Sawtooth Software. Research Paper Series*.
- Ortner, S. B., & Whitehead, H. (Eds.). (1981). Sexual meanings: The cultural construction of gender and sexuality. *CUP Archive*.
- Paivio, A. (1971). Imagery and Language. In S. J. Segal (Ed.), *Imagery: Current Cognitive Approaches* (pp. 7-32). New York, NY: Academic Press.
- Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 45(3), 255-287.
- Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, 21(6), 425-433.
- Parhi, D. R. (2018). Advancement in navigational path planning of robots using various artificial and computing techniques. *Int Rob Auto Journal*, 4(2), 133-136.
- Parker, P. (2016). *Routledge Revivals: Literary Fat Ladies (1987): Rhetoric, Gender, Property*. London: Routledge.
- Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8(12), 976-987.
- Pedreshi, D., Ruggieri, S., & Turini, F. (2008, August). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 560-568). ACM.
- Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
- Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S., Kanwisher, N., ... & Sudarsky, S. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature*, 9(963), 1-13.
- Pessin, J. (1932). The effect of similar and dissimilar conditions upon learning and relearning. *Journal of Experimental Psychology*, 15(4), 427-435.
- Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., Spivey, M., & McRae, K. (2011). The mechanics of embodiment: A dialog on

- embodiment and computational modeling. *Frontiers in Psychology*, 2, 1-5.
- Pfeifer, R., & Scheier, C. (2001). *Understanding intelligence*. Cambridge, MA: MIT Press.
- Postle, B. R. (2015). The cognitive neuroscience of visual short-term memory. *Current Opinion in Behavioral Sciences*, 1, 40-46.
- Pulvermüller, F., Cooper-Pye, E., Dine, C., Hauk, O., Nestor, P. J., & Patterson, K. (2010). The word processing deficit in semantic dementia: all categories are equal, but some categories are more equal than others. *Journal of Cognitive Neuroscience*, 22(9), 2027-2041.
- Pyle, D., & San Jose, C. (2015). An executive's guide to machine learning. *McKinsey Quarterly*, (3), 44-53.
- Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12(5), 410-430.
- Quine, W. V., & Ullian, J. S. (1970). *The web of belief*. New York, NY: McGraw-Hill.
- Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J. G. (2013). Non-graphical solutions for Cattell's scree test. *Methodology*, 9(1), 23-29.
- Ramisa, A., Yan, F., Moreno-Noguer, F., & Mikolajczyk, K. (2018). Breakingnews: Article annotation by image and text processing. *IEEE transactions on pattern analysis and machine intelligence*, 40(5), 1072-1085.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336), 846-850.
- Recchia, G., & Jones, M. (2012). The semantic richness of abstract concepts. *Frontiers in Human Neuroscience*, 6, 315, 1-16.
- Reitman, W. R., Grove, R. B., & Shoup, R. G. (1964). Argus: An information-processing model of thinking. *Behavioral Science*, 9(3), 270-281.
- Reuben, E., Sapienza, P., & Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 201314788.
- Richards, N. M., & King, J. H. (2014). Big data ethics. *Wake Forest L. Rev.*, 49, 393-432.
- Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. *ArXiv Preprint ArXiv:1706.08606*.
- Roediger, H. L. (1980). Memory metaphors in cognitive psychology. *Memory & Cognition*, 8(3), 231-246.
- Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA: MIT Press.

- Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3), 192-233.
- Rosch, E. (1977). Human categorization. *Studies in Cross-cultural Psychology*, 1, 1-49.
- Rosch, E. (1999). Principles of categorization. *Concepts: Core Readings*, 1, 189-206.
- Ross, K., & Carter, C. (2011). Women and news: A long and winding road. *Media, Culture & Society*, 33(8), 1148-1165.
- Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), 2323-2326.
- Rudman, L. A., & Kilianski, S. E. (2000). Implicit and explicit attitudes toward female authority. *Personality and Social Psychology Bulletin*, 26(11), 1315-1328.
- Rudman, L. A., & Phelan, J. E. (2010). The effect of priming gender roles on women's implicit gender beliefs and career aspirations. *Social Psychology*, 41(3), 192-202.
- Rudman, L., Greenwald, A. G., & McGhee, D. E. (1996, October). Powerful women, warm men? Implicit associations among gender, potency, and nurturance. In *Meeting of the Society of Experimental Social Psychology*, Sturbridge, MA.
- Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks* (pp. 405-420). San Diego, CA: Academic Press.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536.
- Sanders, C. K. (2015). Economic abuse in the lives of women abused by an intimate partner: A qualitative study. *Violence Against Women*, 21(1), 3-29.
- Sanford, A. J., Garrod, S., & Boyle, J. M. (1977). An independence of mechanism in the origins of reading and classification-related semantic distance effects. *Memory & Cognition*, 5(2), 214-220.
- Saussure, F. (1916). Nature of the Linguistic Sign. *Cours de linguistique generale*. In D. Richter (Ed.), *Critical Tradition: Classic Texts and Contemporary Trends* (pp. 832-835). Boston: Bedford.
- Schacter, D.L., Bowers, J., Booker, J. (1989). Intention, awareness and implicit memory: the retrieval intentionality criterion. In S. Lewandowsky, J.C. Dunn, & K. Kirsner (Eds.), *Implicit Memory: Theoretical Issues* (pp. 47-65). Hillsdale, NJ: Lawrence Erlbaum Associates.

- Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, 18(2), 173-187.
- Schmittmann, V. D., Cramer, A. O., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, 31(1), 43-53.
- Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, 34(8), 1096-1109.
- Schoenherr, T., Ellram, L. M., & Tate, W. L. (2015). A note on the use of survey research firms to enable empirical data collection. *Journal of Business Logistics*, 36(3), 288-300.
- Schwanenflugel, P. J. (1991). Contextual constraint and lexical processing. *Advances in Psychology*, 77, 23-45.
- Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(1), 82-102.
- Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, 27(5), 499-520.
- Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science*, 5(4), 195-200.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-424.
- Sejnowski, T. J. (1981). Skeleton filters in the brain. In G.E. Hinton, & J.A. Anderson (Eds.). *Parallel models of associative memory* (pp. 189-212). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Sha, L., Haxby, J. V., Abdi, H., Guntupalli, J. S., Oosterhof, N. N., Halchenko, Y. O., & Connolly, A. C. (2015). The animacy continuum in the human ventral vision pathway. *Journal of Cognitive Neuroscience*, 27(4), 665-678.
- Shallice, T. (1988). *From neuropsychology to mental structure*. New York, NY: Cambridge University Press.
- Shallice, T., & Cooper, R. (2011). *The organisation of mind*. New York: Oxford University Press.
- Shallice, T., & Cooper, R. (2013). Is there a semantic system for abstract words?. *Frontiers in Human Neuroscience*, 7(175), 1-10.

- Shallice, T., Warrington, E. K., & McCarthy, R. (1983). Reading without semantics. *The Quarterly Journal of Experimental Psychology Section A*, 35(1), 111-138.
- Shankar, D., Narumanchi, S., Ananya, H. A., Kompalli, P., & Chaudhury, K. (2017). Deep learning based large scale visual recommendation and search for e-commerce. *ArXiv Preprint ArXiv:1703.02344*.
- Shapiro, L. (2011). *New problems of philosophy. Embodied cognition*. New York, NY: Routledge.
- Shea, N. (2018). Metacognition and abstract concepts. *Phil. Trans. R. Soc. B*, 373(1752), 20170133.
- Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information. *ArXiv Preprint ArXiv:1703.00810*.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484-489.
- Sinha, N. (1979). *Sāṃkhya Philosophy*. Allahabad, India: Panini Office.
- Small, S. L., Hart, J., Nguyen, T., & Gordon, B. (1995). Distributed representations of semantic knowledge in the brain. *Brain*, 118(2), 441-453.
- Smith, M. C., Theodor, L., & Franklin, P. E. (1983). The relationship between contextual facilitation and depth of processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(4), 697-712.
- Smolensky, P. (1988). Connectionism and the language of thought. In B. Loewer & G. Rey (Eds.), *Meaning in Mind: Fodor and his Critics* (pp. 201-227). Oxford: Basil Blackwell.
- Sousa, D. (Ed.). (2010). *Mind, brain, and education: Neuroscience implications for the classroom*. Bloomington, IN: Solution Tree Press.
- Sowa, J. F. (1976). Conceptual graphs for a data base interface. *IBM Journal of Research and Development*, 20(4), 336-357.
- Sowa, J. F. (1979, June). Semantics of conceptual graphs. In *Proceedings of the 17th annual meeting on Association for Computational Linguistics* (pp. 39-44). Association for Computational Linguistics.
- Sporns, O. (2010). *Networks of the Brain*. Cambridge, MA: MIT Press.
- Sporns, O. (2011). The human connectome: A complex network. *Annals of the New York Academy of Sciences*, 1224(1), 109-125.
- Sporns, O., & Zwi, J. D. (2004). The small world of the cerebral cortex. *Neuroinformatics*, 2(2), 145-162.
- Stewart, J., Stewart, J. R., Gapenne, O., & Di Paolo, E. A. (Eds.). (2010). *Enaction: Toward a new paradigm for cognitive science*. Cambridge, MA: MIT Press.

- Stramandinoli, F., Marocco, D., & Cangelosi, A. (2017). Making sense of words: A robotic model for language abstraction. *Autonomous Robots*, 41(2), 367-383.
- Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., & Mitchell, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62(1), 451-463.
- Sun, R. (Ed.). (2008). *The Cambridge Handbook of Computational Psychology*. New York, NY: Cambridge University Press.
- Svare, H. (2006). *Body and practice in Kant* (Vol. 6). Berlin: Springer-Verlag.
- Sweeney, L. (2013). Discrimination in online ad delivery. *ArXiv Preprint ArXiv:1301.6822*.
- Talmy, L. (1996). Fictive motion in language and “ception.” In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (pp. 211–275). Cambridge, MA: MIT Press.
- Tarski, A. (1935). Zur Grundlegung der Boole'schen Algebra. I. *Fundamenta mathematicae*, 1(24), 177-198.
- Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D. D. L., ... & Lillicrap, T. (2018). DeepMind Control Suite. *ArXiv Preprint ArXiv:1801.00690*.
- Taylor, G. A., & Juola, J. F. (1974). Priming effects on recognition performance. *Bulletin of the Psychonomic Society*, 3(4), 277-279.
- Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2319-2323.
- Teuber, H. L. (1955). Physiological psychology. *Annual Review of Psychology*, 6(1), 267-296.
- Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision*, 53(2), 169-191.
- Touroutoglou, A., Lindquist, K. A., Dickerson, B. C., & Barrett, L. F. (2015). Intrinsic connectivity in the human brain does not reveal networks for ‘basic’ emotions. *Social Cognitive and Affective Neuroscience*, 10(9), 1257-1265.
- Trafton, J. A., Sorrell, J. T., Holodniy, M., Pierson, H., Link, P., Combs, A., & Israelski, D. (2012). Outcomes associated with a cognitive-behavioral chronic pain management program implemented in three public HIV primary care clinics. *The Journal of Behavioral Health Services & Research*, 39(2), 158-173.
- Tranel, D. (2006). Impaired naming of unique landmarks is associated with left temporal polar damage. *Neuropsychology*, 20(1), 1-10.

- Troche, J., Crutch, S. J., & Reilly, J. (2017). Defining a Conceptual Topography of Word Concreteness: Clustering Properties of Emotion, Sensation, and Magnitude among 750 English Words. *Frontiers in Psychology*, 8(1787), 1-15.
- Troche, J., Crutch, S., & Reilly, J. (2014). Clustering, hierarchical organization, and the topography of abstract and concrete nouns. *Frontiers in Psychology*, 5(360), 1-10.
- Tucker, M., & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception and Performance*, 24(3), 830-846.
- Tulving, E. (1972). Episodic and semantic memory. *Organization of Memory*, 1, 381-403.
- Tulving, E. (1983). Ecphoric processes in episodic memory. *Phil. Trans. R. Soc. Lond. B*, 302(1110), 361-371.
- Turner, M. (1996). *The literary mind: The origins of thought and language*. New York: Oxford University Press.
- Tyler, A., & Evans, V. (2003). *The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition*. New York, NY: Cambridge University Press.
- Vallet, G. T. (2015). Embodied cognition of aging. *Frontiers in Psychology*, 6(463), 1-6.
- Van Dantzig, S., Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2008). Perceptual processing affects conceptual processing. *Cognitive Science*, 32(3), 579-590.
- Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579-2605.
- Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature News*, 514(7524), 550-553.
- Van Orden, G. C., Pennington, B. F., & Stone, G. O. (2001). What do double dissociations prove?. *Cognitive Science*, 25(1), 111-172.
- Varela, F., Thompson, E., & Rosch, E. (1991). *The Embodied Mind*. Cambridge, MA: MIT Press.
- Vigliocco, G., & Vinson, D. P. (2007). Semantic Representation. In M.G. Gaskell, & G.T.M. Altmann (Eds.), *The Oxford Handbook of Psycholinguistics* (pp. 195 – 215). New York: Oxford University Press.
- Vigliocco, G., Kousta, S. T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2013). The Neural Representation of Abstract Words: The Role of Emotion. *Cerebral Cortex*, 24(7), 1767-1777.

- Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, 48(4), 422-488.
- Vigneau, M., Beaucousin, V., Herve, P. Y., Duffau, H., Crivello, F., Houde, O., ... & Tzourio-Mazoyer, N. (2006). Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing. *Neuroimage*, 30(4), 1414-1432.
- Walker, I., & Hulme, C. (1999). Concrete words are easier to recall than abstract words: Evidence for a semantic contribution to short-term serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(5), 1256-1271.
- Walker, P. (2016). Cross-sensory correspondences and symbolism in spoken and written language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9), 1339-1361.
- Walker, P., & Parameswaran, C.R. (2019). Cross-sensory correspondences in language: Vowel sounds can symbolize the felt heaviness of objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(2), 246-252.
- Wang, X. J., Zhang, L., Jing, F., & Ma, W. Y. (2006). Annosearch: Image auto-annotation by search. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (Vol. 2, pp. 1483-1490). IEEE.
- Warrington, E. K. (1975). The selective impairment of semantic memory. *The Quarterly Journal of Experimental Psychology*, 27(4), 635-657.
- Warrington, E. K., & McCarthy, R. (1983). Category specific access dysphasia. *Brain*, 106(4), 859-878.
- Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, 107(3), 829-853.
- Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, 20(2), 158-177.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440-442.
- Watts, F. (2013). Embodied Cognition and Religion. *Zygon*, 48(3), 745-758.
- Wellman, H. M., Harris, P. L., Banerjee, M., & Sinclair, A. (1995). Early understanding of emotion: Evidence from natural language. *Cognition & Emotion*, 9(2-3), 117-149.
- Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, 64(6), 956-972.
- West, R. L. (1996). An application of prefrontal cortex function theory to cognitive aging. *Psychological Bulletin*, 120(2), 272-292.

- Wheeler, B. (2014). AlgDesign: Algorithmic experimental design. *R package version*, 1-1.
- Whitley, R. J., Soong, S. J., Linneman, C., Liu, C., Pazin, G., & Alford, C. A. (1982). Herpes simplex encephalitis: Clinical assessment. *Jama*, 247(3), 317-320.
- Whorf, B. (1956). *Language, Thought, and Reality*, ed. J. B. Carroll (Cambridge, MA: MIT Press).
- Wiemer-Hastings, K., & Xu, X. (2005). Content differences for abstract and concrete concepts. *Cognitive Science*, 29(5), 719-736.
- Willems, R. M., & Francken, J. C. (2012). Embodied cognition: Taking the next step. *Frontiers in Psychology*, 3(582), 1-3.
- Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science*, 322(5901), 606-607.
- Winston, P. H. (1984). *Artificial Intelligence*. Reading, MA: Addison-Wesley.
- Witherby, A. E., & Tauber, S. K. (2017). The concreteness effect on judgments of learning: Evaluating the contributions of fluency and beliefs. *Memory & Cognition*, 45(4), 639-650.
- Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell Publishers.
- Wright, A. H. (1991). Genetic algorithms for real parameter optimization. In J.E.R Gregory (Ed.), *Foundations of Genetic Algorithms* (pp. 205-218). San Francisco, CA: Morgan Kaufmann Publishers.
- Wu, L. L., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica*, 132(2), 173-189.
- Xu, H., Murphy, B., & Fyshe, A. (2016). Brainbench: A brain-image test suite for distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2017-2021).
- Yam, J. Y., & Chow, T. W. (2000). A weight initialization method for improving training speed in feedforward neural network. *Neurocomputing*, 30(1-4), 219-232.
- Yamashita, Y., & Tani, J. (2008). Emergence of Functional Hierarchy in a Multiple Timescale Neural Network Model: A Humanoid Robot Experiment. *PLoS Computational Biology*, 4, 1-47.
- Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review*, 18(4), 742-750.

- Yarmey, A. D., & Thomas, K. A. (1966). Set and word abstractness-concreteness shift in paired-associate learning. *Psychonomic Science*, 5(10), 387-388.
- Yi, Y. (1990). A critical review of consumer satisfaction. *Review of Marketing*, 4(1), 68-123.
- Zafar, B. (2013). College major choice and the gender gap. *Journal of Human Resources*, 48(3), 545-595.
- Zahn, R., Moll, J., Iyengar, V., Huey, E. D., Tierney, M., Krueger, F., & Grafman, J. (2009). Social conceptual impairments in frontotemporal lobar degeneration with right anterior temporal hypometabolism. *Brain*, 132(3), 604-616.
- Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017, July). Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (pp. 2881-2890). IEEE.
- Zook, M., Barocas, S., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., ... & Nelson, A. (2017). Ten simple rules for responsible big data research. *PLOS Computational Biology*, 13(3), 1-10.