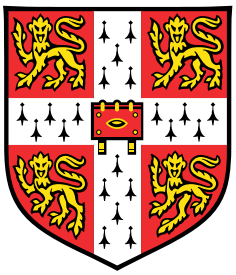
Understanding Semantic Implicit Learning through distributional linguistic patterns

A computational perspective



Dimitrios Alikaniotis

Supervisor: **Dr. J. N. Williams**

Advisors: Dr. Paula Buttery
Dr. Henriëtte Hendriks

Department of Theoretical and Applied Linguistics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Downing College November 2018

It's the most pointless book since "How to learn French" was translated into French.

Mr. Edmund Blackadder

I would like to dedicate this thesis to my loving parents …

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 80,000 words including appendices, footnotes, tables, and equations but excluding bibliographies and has fewer than 150 figures.

Dimitrios Alikaniotis
November 2018

Acknowledgements

I am very fortunate to have many people to thank.

My advisor John N. Williams, for the freedom to go where my heart and mind led me, and for the guidance that kept me from squandering that freedom.

One major tenet of this PhD thesis is that words acquire their meaning by the environments they appear in; I have found this to be particularly true of PhD students as well. As such I would like to thank my friends and fellow group mates Christian Bentz, Andrew Caines, Connor Quinn, Carla Pastorino, and Giulia Bovolenta. I would also like to thank Elspeth Wilson, Nikola Vukovic, Tina Paciorek, Yan Tao, as well as the rest of the regulars of T-R13 for the time I spent there.

I would also like to thank other staff members of the Research Centre for English and Applied Linguistics (now Department of Theoretical and Applied Linguistics), especially Drs. Paula Buttery, Dora Alexpoulou, and Napoleon Katsos for their friendship, encouragement and practical advice. Together with the administrative staff as well as my fellow PhD students and MPhil students, they all contributed to creating a very supportive atmosphere in the department. From the Computer Laboratory, I would like to thank the members of the NLIP group for their advice and comments.

I have also been fortunate to have met and received support from some of the leading figures in the field who were happy to listen to my ideas and provide some helpful comments. Amongst others, I would like to thank Sebastian Padó, Harald R. Baayen, Chris Westbury, Michael Zock, Martijn Wieling, Nicholas Turk-Browne, Martin Chodorow and Marta Kutas.

My gratitude extends to those people who made it possible for me to be here. Professor Christoforos Charalambakis, Drs. M. Kakridi and S. Varlokosta, as well as Prof. P. Mastrodimitris, have all been there when I needed them most. I am also particularly grateful to the Onassis Foundation for which without their support over these three years none of this would have been possible.

Τέλος, εκείνοι στους οποίους αφιερώνεται αυτή η διατριβή. Στους γονείς μου, Διονύσιο και Φωτεινή, καθώς και στον αδερφό μου Παναγιώτη και τη Δανάη για την ανιδιοτελή αγάπη, την αμέριστη συμπαράσταση και κάθε είδους υποστήριξη που δείχνουν σε κάθε μου βήμα και γιατί μου έμαθαν ότι καὶ ἄν ἄνους μὲν ἴω, τοῖς φίλοις δ᾽ ὀρθῶς φίλος …

Abstract

The research presented in this PhD dissertation provides a computational perspective on Semantic Implicit Learning (SIL). It puts forward the idea that SIL does not depend on semantic knowledge as classically conceived but upon semantic-like knowledge gained through distributional analysis of massive linguistic input. Using methods borrowed from the machine learning and artificial intelligence literature we construct computational models, which can simulate the performance observed during behavioural tasks of semantic implicit learning in a human-like way. We link this methodology to the current literature on implicit learning, arguing that this behaviour is a necessary by-product of efficient language processing.

- *Chapter 1* introduces the computational problem posed by implicit learning in general, and semantic implicit learning, in particular, as well as the computational framework used to tackle them.
- *Chapter 2* introduces distributional semantics models as a way to learn semantic-like representations from exposure to linguistic input.
- *Chapter 3* reports two studies on large datasets of semantic priming which seek to identify the computational model of semantic knowledge that best fits the data under conditions that resemble SIL tasks. We find that a model which acquires semantic-like knowledge gained through distributional analysis of massive linguistic input provides the best fit to the data.
- *Chapter 4* generalises the results of the previous two studies by looking at the performance of the same models in languages other than English.
- *Chapter 5* applies the results of the two previous Chapters on eight datasets of semantic implicit learning. Crucially, these datasets use various semantic manipulations and speakers of different L1s enabling us to test the predictions of different models of semantics.
- *Chapter 6* examines more closely two assumptions which we have taken for granted throughout this thesis. Firstly, we test whether a simpler model based on phonological

information can explain the generalisation patterns observed in the tasks. Secondly, we examine whether our definition of the computational problem in Chapter 5 is reasonable.

- *Chapter 7* summarises and discusses the implications for implicit language learning and computational models of cognition. Furthermore, we offer one more study that seeks to bridge the literature on distributional models of semantics to 'deeper' models of semantics by learning semantic relations.

There are two main contributions of this dissertation to the general field of implicit learning research. Firstly, we highlight the superiority of distributional models of semantics in modelling unconscious semantic knowledge. Secondly, we question whether 'deep' semantic knowledge is needed to achieve above chance performance in SIL tasks. We show how a simple model that learns through distributional analysis of the patterns found in the linguistic input can match the behavioural results in different languages. Furthermore, we link these models to more general problems faced in psycholinguistics such as language processing and learning of semantic relations.

Contents

List of Figures **xix**

List of Tables **xxi**

1 Introduction **1**
1.1 Implicit Learning . . . 2
1.2 The computational problem of implicit learning . . . 4
1.3 The computational framework . . . 10
1.3.1 Representations . . . 18
1.4 Implicit Language Learning . . . 23
1.5 Semantic Implicit Learning . . . 25

2 Distributional Semantics **35**
2.1 Introduction . . . 35
2.2 Vector-Space Models . . . 37
2.2.1 Syntagmatic Models . . . 39
2.2.2 Paradigmatic Models . . . 44
2.2.3 BEAGLE . . . 46
2.3 Predictive models . . . 48
2.3.1 Neural Embeddings . . . 48
2.3.2 Recurrent Neural Embeddings . . . 53
2.4 General considerations . . . 54
2.4.1 Multiword Expressions . . . 54
2.4.2 Corpus choice and parameter spaces . . . 55

3 Discovering the unconscious representations **57**
3.1 Introduction . . . 57
3.2 Semantic Priming . . . 59
3.2.1 Prior Work . . . 60

3.2.2 Impact on models of semantic priming . . . 62
3.3 Method . . . 64
3.3.1 Model Selection . . . 64
3.3.2 Baselines . . . 66
3.4 Study 1: Priming Effects in the Semantic Priming Project . . . 69
3.4.1 Dataset . . . 69
3.4.2 Splitting the dataset . . . 70
3.4.3 Results and discussion . . . 71
3.5 Study 2: Mediated Semantic Priming . . . 80
3.5.1 Method . . . 80
3.5.2 Dataset . . . 82
3.5.3 Results and discussion . . . 82
3.6 Discussion for Studies 1 and 2 . . . 87
3.6.1 Reliability Issues . . . 88

4 Cross-linguistic exploration of the unconscious representations 95
4.1 Introduction . . . 95
4.2 Semantic Neighbourhood Density . . . 98
4.2.1 Prior work . . . 98
4.3 Method . . . 99
4.3.1 Datasets . . . 101
4.3.2 Corpora . . . 102
4.3.3 Baselines . . . 104
4.4 Study 3: Semantic density effects in English . . . 106
4.4.1 Method . . . 106
4.4.2 Results and discussion . . . 107
4.5 Study 4: Semantic density effects in other languages . . . 112
4.5.1 Method . . . 112
4.5.2 Results and discussion . . . 112
4.6 Discussion for Studies 3 and 4 . . . 113
4.7 General Discussion for Studies 1–4 . . . 115

5 Distributional Semantics Approach to Implicit Language Learning 119
5.1 Introduction . . . 119
5.2 Computational overview of the tasks . . . 119
5.3 Method . . . 125
5.3.1 Feed-forward neural network . . . 125

5.3.2 Evaluation . . . 129
5.3.3 Visualising the spaces . . . 130
5.3.4 A note on non-distributional representations . . . 131
5.4 Study 5: Animate / Inanimate . . . 133
5.4.1 Introduction . . . 133
5.4.2 Materials . . . 135
5.4.3 Results and discussion . . . 136
5.5 Study 6: Abstract / Concrete . . . 139
5.5.1 Introduction . . . 139
5.5.2 Materials . . . 142
5.5.3 Results and discussion . . . 143
5.6 Study 7: Perceptual features . . . 146
5.6.1 Introduction . . . 146
5.6.2 Materials . . . 148
5.6.3 Results and discussion . . . 148
5.7 Study 8: Language-Specific distributional cues . . . 150
5.7.1 Introduction . . . 150
5.7.2 Materials . . . 152
5.7.3 Results and discussion . . . 152
5.8 Discussion for Studies 5–8 . . . 156
5.8.1 A theory of Semantic Implicit Learning . . . 157
5.8.2 Linguistic usage and semantic features . . . 161
5.8.3 Computational considerations . . . 169

6 Checking the assumptions **175**
6.1 Introduction . . . 175
6.2 Study 9: Is Semantic Implicit Learning really semantic? . . . 175
6.2.1 Introduction . . . 175
6.2.2 Materials and methods . . . 179
6.2.3 Results . . . 179
6.2.4 Discussion . . . 182
6.3 Study 10: Comparing forward and backward probabilities . . . 184
6.3.1 Introduction . . . 184
6.3.2 Materials and methods . . . 187
6.3.3 Results . . . 189
6.3.4 Discussion . . . 192

7 General Discussion **197**
7.1 Summary of findings . . . 197
7.2 Study 11: Beyond co-occurences? . . . 199
7.2.1 Introduction . . . 199
7.2.2 Materials and Methods . . . 200
7.2.3 Results and discussion . . . 201
7.2.4 General Discussion . . . 208
7.3 Implications of current findings and future research . . . 210

Abbreviations **213**

Notation **217**

Bibliography **219**

Appendix A Details for the in-text studies **253**
A.1 Generating experimental stimuli from the FSG . . . 253
A.2 Generating the data for Fig. 1.3 . . . 254
A.3 Clockwork orange . . . 256
A.4 Semantic Priming Project . . . 257
A.5 Evaluating dimensionality reduction in WordNet . . . 258
A.6 Training Italian neural embeddings . . . 260
A.7 Phonological correlates of semantics . . . 260

Appendix B Simulation Details **267**
B.1 Initialisation of the weights . . . 267
B.2 Bayesian Optimiser parameter spaces . . . 267
B.3 Computing the semantic neighbourhoods . . . 269
B.4 Phonological Cues . . . 270

Appendix C Dataset Details **273**
C.1 Animate / Inanimate . . . 273
C.2 Abstract / Concrete . . . 275
C.2.1 High similarity . . . 275
C.2.2 Low similarity . . . 277
C.3 Perceptual features . . . 279
C.4 Language-Specific distributional cues . . . 282
C.4.1 Mandarin Chinese . . . 282

C.4.2 English 285
C.5 Stimuli from Saalbach & Imai (2007) 289

List of Figures

1.1 The finite-state grammar used in Reber (1967) 4
1.2 Space requirements of the different memorisation methods 7
1.3 Example of an unlearnable input and its solution 16
1.4 WordNet hierarchy for the synset `dog.n.01` 23
1.5 The semantically-driven determiner system used in Williams (2005) 28
1.6 A single trial from Leung & Williams (2014) 31
1.7 The original MNPQ paradigm 32

2.1 Two-dimensional projection of distributed semantic representations 37
2.2 Neural Embeddings Learning 50
2.3 Neural Embeddings Learning by Negative Sampling 51

3.1 Study 1: Semantic relations in the Semantic Priming Project (SPP) 77

5.1 Architecture of the connetionist model 126
5.2 Study 5: Two-dimensional projection of the stimuli used in the 'Animate / Inanimate' simulations 137
5.3 Study 5: Generalisation gradients and hidden layer activations for the 'Animate / Inanimate' simulations 138
5.4 Study 6: Two-dimensional projection of the stimuli used in the 'Abstract / Concrete' simulations 144
5.5 Study 6: Point-estimates from the two datasets used in the 'Abstract / Concrete' simulations 145
5.6 Study 6: Two-dimensional projection of the activation of the hidden layers given the training stimuli used in the two 'Abstract / Concrete' simulations . 146
5.7 Study 7: Two-dimensional projection of the stimuli used in the 'Perceptual features' simulations 149
5.8 Study 8: Two-dimensional projection of the stimuli used in the 'Language-Specific distributional cues' simulations 153

5.9 Study 8: Generalisation gradients for the 'Language-Specific distributional cues' simulations . . . 154
5.10 Correlation between observed behavioural performance on the Two alternative forced choice task (2AFC) tasks and corresponding model predictions. . 157
5.11 Study 8: Results from five simulated learners in the high-similarity 'Abstract / Concrete' simulation . . . 162
5.12 Two-dimensional projection of 300 Italian nouns by colour-coded by article 163
5.13 Two-dimensional projection of 8000 English nouns colour-coded for concreteness . . . 167
5.14 Two-dimensional projection of 217 words which include either `is_large` or `is_small` as features in the McRae norms (McRae, Cree, Seidenberg & McNorgan, 2005) . . . 170

6.1 Study 9: Two-dimensional projection of the phonological space of the nouns used in the semantic implicit learning experiments . . . 182
6.2 Study 10: Forward and backward entropy rates per thousand words on ten languages from the The Europarl parallel corpus (EUROPARL) corpus. . . . 190
6.3 Study 10: Three way interaction between `tp` × `type` × `article` . . . 194

7.1 Study 11: Extending the architecture of Rumelhart & Todd (1993) . . . 201
7.2 Study 11: Results of the simulations on the dataset . . . 202

A.1 Feedforward neural network projecting 2D input in 3D space . . . 255
A.2 Parsed sentence from the Clockwork Orange . . . 256
A.3 Comparison of the correlation coefficients between WordNet vectorised representations and human similarity ratings . . . 259
A.4 Phonological features predictive of *animacy* and *concreteness* . . . 263
A.5 Results of the semantic classification according to phonological features . . . 264

List of Tables

1.1 Internal representations for the memorisation models. 9

2.1 Example Term × Document Matrix prior to normalisation 40
2.2 Example Term × Document Matrix after *tf-idf* 41
2.3 Common Distributional Semantics Model (DSM) normalisation techniques . 41
2.4 Common vector similarity measures 42
2.5 Example Term × Term co-occurrence matrix 45

3.1 Study 1: Summary of the best performing models on the validation and testing sets 72
3.2 Study 1: Pearson correlation coefficients between the similarity measure of each DSM and the priming effects in the naming task of the SPP 73
3.3 Study 1: Best parameter sets for each of the DSMs 75
3.4 Study 1: Examples of semantic relations 78
3.5 Study 1: Average human priming effects and model predictions aggregated by semantic relation 79
3.6 Study 2: Pearson correlation coefficients between priming effects and model predictions aggregated by relation 81
3.7 Study 2: Studies included in Jones, Kintsch & Mewhort (2006) 82
3.8 Study 2: Results from Balota & Lorch (1986) 83
3.9 Study 2: Results from McKoon & Ratcliff (1992) 86
3.10 Average Mutual Information between the word pairs in McKoon & Ratcliff (1992) and Balota & Lorch (1986). 93

4.1 Summary of the corpora described in §4.3.2 and used in the simulations reported in §4.5. 105
4.2 Study 3: Summary of the best performing models on the SPP 108
4.4 Study 3: Summary of the best performing models on the British Language Project 110

4.5 Study 4: Best parameter sets for each of the DSMs . . . 113
4.6 Study 4: Summary of the best performing models on the validation and testing sets . . . 114

5.1 Summary of the simulated semantic implicit learning experiments. . . . 132
5.2 Study 6: Examples of high- and low-similarity stimuli from Paciorek & Williams (2015) . . . 142
5.3 The Chinese classifiers used in the clustering simulations along with the number of words using that classifier . . . 164
5.4 Study 8: The ten most frequent features given as responses in the McRae norms 168

6.1 Study 9: Example phonological feature vectors . . . 180
6.2 Study 9: By-feature means for the stimuli used in the semantic implicit learning experiments . . . 181
6.3 Study 10: Information on the languages used from the EUROPARL corpus. . . 187
6.4 Study 10: Summaries of the Ordinary Least Squares (OLS) model and the Linear mixed-effects regression model (LMM) model predicting the surprisal values . . . 193

7.1 Study 11: Model predictions on a set of novel words . . . 206
7.2 Study 11: Semantic categories used and the corresponding WordNet synsets . 207
7.3 Study 11: Model performance on two semantic categories . . . 208

A.1 Transitional probabilities from the Finite-state grammar (FSG) depicted in Fig. 1.1 . . . 254
A.2 Summary of the baseline model used in §3.4 . . . 258
A.3 WordNet synsets used in the simulations . . . 261
A.4 Performance of the trained classifier on the datasets used in the behavioural experiments . . . 262

B.1 Parameter spaces for each of the models tested in §§ 3.4 and 4.5 . . . 268
B.2 Vowel height and vowel position values used to generate the phonological representations . . . 270
B.3 Description of the features used in §6.3 . . . 271
B.4 Division of phonemes in the ARPABET format . . . 272

C.1 Training materials for the 'Animate / Inanimate' simulations . . . 274
C.2 Testing materials for the 'Animate / Inanimate' simulations . . . 275
C.3 Training materials for the 'Abstract / Concrete (High similarity)' simulations 275

C.4 Testing materials for the 'Abstract / Concrete (High similarity)' simulations . 276
C.5 Training materials for the 'Abstract / Concrete (Low similarity)' simulations 277
C.6 Testing materials for the 'Abstract / Concrete (Low similarity)' simulations . 278
C.7 Training materials for the 'Perceptual features' simulations 279
C.8 Testing materials for the 'Perceptual features' simulations 281
C.9 Training materials for the 'Language-Specific distributional cues (Mandarin Chinese)' simulations 282
C.10 Testing materials for the 'Language-Specific distributional cues (Mandarin Chinese)' simulations 284
C.11 Training materials for the 'Language-Specific distributional cues (English)' simulations 285
C.12 Testing materials for the 'Language-Specific distributional cues (English)' simulations 287
C.13 Stimuli used in Saalbach & Imai (2007) 289

Chapter 1

Introduction

> When we're learning to see, nobody's telling us what the right answers are — we just look. Every so often, your mother says "that's a dog", but that's very little information. You'd be lucky if you got a few bits of information — even one bit per second — that way. The brain's visual system has 10^{14} neural connections. And you only live for 10^9 seconds.[1] So it's no use learning one bit per second. You need more like 10^5 bits per second. And there's only one place you can get that much information: from the input itself.
>
> Geoffrey Hinton, 1996 (quoted in Gorder, 2006)

Generally speaking, texts on implicit learning introduce the topic with some complex computation that the human mind can perform accurately, the steps of which, however, we are unable to verbalize. Common examples range from *intuitive physics* and how we learn to steer a bike (Eysenck, 2008) or judge the trajectory of a ball (Reed, McLeod & Dienes, 2010) to *intuitive psychology* and *social skills* (Lieberman, 2000) to higher levels of human cognition such as how we *learn and process language* (Williams, 2009). Perhaps the ubiquitousness of the phenomenon and the relative ease by which humans learn certain complex skills can explain the persistence of the community in reusing such examples. Calculating, for instance, the trajectory of a projectile is a convoluted issue requiring knowledge of the initial speed and the angle at which the projectile was launched as well as taking into account air resistance and gravitational pull. Implicit learning (IL) is the process of learning complex information from the statistical regularities provided by the environment without being aware that we are doing so. For example, a five-year-old child does not need to understand Newton's second Law of

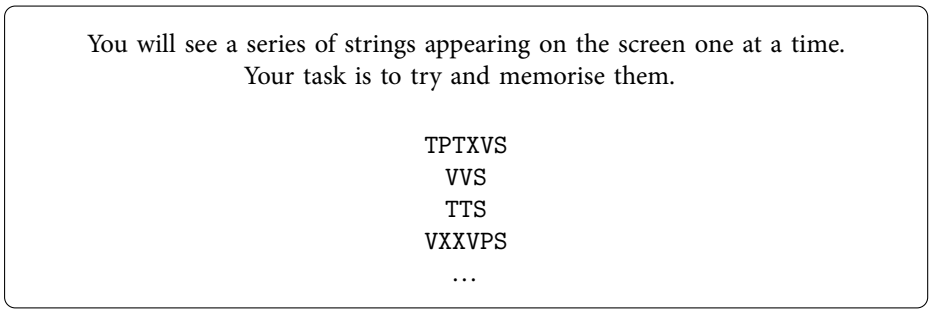
[1] The figures in this quote are not to be taken literally as their point is to show the differences between the orders of magnitude. 10^9 seconds is around 31.71 years. Instead, as of 2015 the worldwide life expectancy is about 71 years (min 50.1) (World Health Organization, 2016) making the figure closer to 2.24×10^9.

Motion which is necessary to compute the trajectory of a ball but can use simple heuristics to catch a ball.

The present thesis aspires at providing a computational framework within which we can explore the interaction between the unconscious extraction of statistical regularities and language learning. Following usage-based theories of language acquisition (Tomasello, 2003), we take as our starting point that constant exposure to a linguistic environment containing rich statistical information biases language processing in a way that leads us to learn languages more efficiently. The thesis is organised as follows; Chapter 1 introduces the topic of implicit learning and the computational problem that it poses along with the computational framework we will be using. Chapter 2 explores the nature of the information contained in the linguistic input and how we can extract it. Chapters 3 and 4 look at behavioural experiments exploring our internal linguistic representations and seek to find the best computational description of them. In addition, we also look at whether our results hold cross-linguistically. We then move on to examine how these unconscious linguistic representations bias our learning mechanisms enabling us to learn information beyond what is given (Chapter 5). Chapter 6 checks some of the assumptions put forward in the previous chapters and Chapter 7 provides a general discussion of the previous results, explores how exposure to language might lead to structured linguistic knowledge and provides ideas for future work.

1.1 Implicit Learning

Much like every statistics textbook must contain a coin-flipping example because this is the most intuitive way to introduce rudimentary notions such as Bernoulli processes and the binomial distribution, any text on implicit learning *must* contain an example of a task the brain does without us being able to explain why. For this we will follow the classic examples by Reber (1967) (also Reber, 1989); using any off-the-shelf psychophysics library, we can write a series of computer programs which continuously emit nonsense strings to the user.



As per the instructions, the task of the participant is to merely memorise the presented strings as we did not mention that they might subsequently be tested. After some training, we inform the participants that an underlying grammar generated these strings and that their task is to judge which of a set of new strings were generated by the same grammar:

> Were these strings generated by the same grammar?
>
> TPTS
>
> TPTXPS

Figure 1.1 shows the finite-state machine (i.e., the 'grammar') used to produce all the above strings. Under this grammar, the string TPTS should be grammatical as there is an unbreakable path from the initial (s_0) to the accepting (s'_0) states while for TPTXPS there is none. What makes implicit learning particularly interesting is that in the subsequent *generalisation* phase participants can classify the novel strings as either grammatical or ungrammatical *without being able to verbalise the structure of the underlying grammar*.

Subsequent research on implicit learning has shed light on a number of issues such as whether participants split the input sequences to smaller *chunks* (Perruchet & Pacteau, 1990; Perruchet, Vinter, Pacteau & Gallego, 2002; Servan-Schreiber & Anderson, 1990), use ad hoc *mini*-rules (e.g., *sequences start with either* 'T' *or* 'V'), use analogical or rule-like reasoning (McAndrews & Moscovitch, 1985; Opitz & Hofmann, 2015), are affected by the complexity of the grammar (Domangue, Mathews, Sun, Roussel & Guidry, 2004; Mathews, Buss, Chinn & Stanley, 1988; Mathews, Buss, Stanley, Blanchard-Fields, Cho & Druhan, 1989; Stanley, Mathews, Buss & Kotler-Cope, 1989), are affected by the frequencies of smaller chunks within the sequences (Knowlton & Squire, 1994; Meulemans & der Linden, 1997), whether memory impairments hinder this kind of learning (Abrams & Reber, 1988; Knowlton, Ramus & Squire, 1992), whether input modality plays a role or whether participants could transfer their knowledge to different letter-sets (Altmann, Dienes & Goode, 1995) or whether sleep consolidation plays any role (Nieuwenhuis, Folia, Forkstam, Jensen & Petersson, 2013).

The above literature shows that we know a good deal of the process and the limits of learning artificial grammars (as in Fig. 1.1) implicitly as well as some of the computational mechanisms involved. This research has shown how domain-general learning mechanisms involved in skill learning (Sun, 1997) play a role in IL (Pacton, Sobaco & Perruchet, 2015; Perruchet & Pacton, 2006; Perruchet & Rey, 2005; Perruchet & Vinter, 2002). In what follows, we will introduce an account of the computational problem posed by implicit learning and the tools we will be using to explain it.

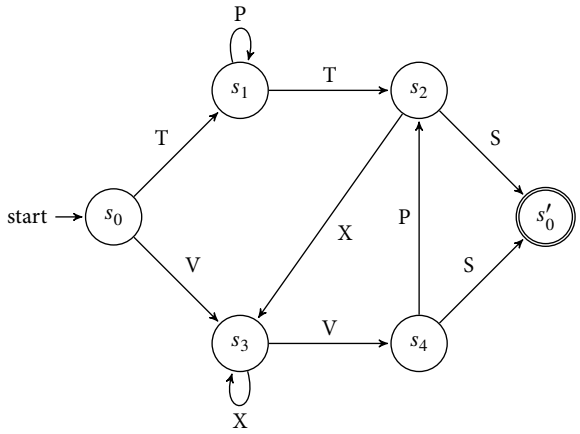


Figure 1.1 The finite-state grammar used in the implicit learning experiments done by Reber (1967). During training, the grammar generates strings starting from the left (s_0) following a random path until it reaches the accepting state on the right (s'_0). Self-connections denote potential loops which yield strings like $\mathtt{TP}^n\mathtt{TS}$ (where n is the number of repetitions). During testing the participants come across strings either generated by the grammar or random strings with the same alphabet. A string can be classified as grammatical if there is an unbreakable path from the initial to the accepting state. Otherwise, it is considered ungrammatical.

1.2 The computational problem of implicit learning

Implicit learning is acquisition of knowledge about the underlying structure of a complex stimulus environment by a process which takes place naturally, simply and without conscious operations (Ellis, 1994). The situation we outlined above meets all the above criteria for implicit learning. The *complex stimulus environment* the participants come across is the finite-state grammar, *knowledge about the underlying structure* is demonstrated by the above chance performance in the subsequent generalisation tasks, and the lack of *conscious operations* can be seen from various post-experimental procedures which assess conscious knowledge of the rules as well as the fact that participants were instructed to *memorise* the strings instead of actively figuring out any rules. The question we pose at this point is the following; what computational mechanisms can turn mere memorisation to uncovering a highly complex structure?

Consider again the nonsensical strings produced by the finite-state grammar in Fig. 1.1. If the task were to recall only the strings that were displayed during the training phase and were generated by the FSG as opposed to random strings (as in, e.g., Miller, 1958) we could argue that the learner would only need to store the patterns in a temporary buffer during the familiarisation phase and then retrieve them during testing. Mere memorization, however,

has at least two problems; firstly, from a computational[2] point of view it has intensive space requirements $\Theta(n)$,[3] where n is the number of training patterns (Cleeremans, 1997, for a similar point on *serial reaction time* tasks), and secondly, it could not explain above chance performance in the generalisation task. A solution to this problem would be to assume that during familiarisation participants do not memorise the patterns but *abstract* characteristics of the patterns that enable them to *learn* something about them instead of only storing them.

The nature of this *abstraction* process has been a widely debated issue, and we have already mentioned proposals ranging from participants learning mini-rules (e.g., *each string starts with either 'V' or 'T'*), to learning to recognise smaller units (*chunks)*. Participants can then use that knowledge to judge the grammaticality of novel instances (e.g., *a string starting with 'V' might be legal* while *a string containing the bigram 'VT' cannot be legal*). While numerous computational models have been proposed to account for the performance in such tasks, it is beyond the scope of the present thesis to explore them in detail (see Cleeremans & Dienes, 2008 for a computational overview). However, it would be pertinent to this introduction to examine how they explain the *dissociation* between what the mind *does* (i.e., *abstraction*) against what is *trying* to do (i.e., *memorisation*). For instance, the computer programs we wrote before asked participants to memorise the sequences discouraging them from seeking out the rules generating these strings. The participants, nevertheless, can classify novel instances without explicit knowledge of the underlying system, pointing to the direction that either they approach the task differently, or, that the memorisation algorithm enables them to learn something beyond the given patterns without attempting to do so.

We argue that this dissociation comes from the specific algorithmic implementation the mind is using to memorise the patterns more efficiently. In other words, the computational problem the mind is solving is still *memorisation*, but the specific memorisation algorithm used gives rise to learning more than the presented sequences. This clear distinction between the two levels of description, the computational problem and its algorithmic implementation, comes from Marr (1982)[4] (also see Anderson, 1990 for an application in cognitive psychology) who argues that any psychological theory should start from the computational level of description

[2] As it will become clearer later in this chapter, we use the terms *computational* and *algorithmic* in a purely Marr (1982) sense. The *computational* issue here is taken to define an abstract problem that needs to be solved, while the algorithmic is a step-by-step solution to the problem. Note that the same computational problem can have will have multiple algorithmic solutions.

[3] Notationally, we use Θ to denote that a function grows as fast as the argument (here n) and O to denote a function grows *no faster* than its argument. In other words, $\Theta(n)$ in this context should read as 'the space requirements grow as fast as n'. An example of O would be what we call the reduced method of storing later which has space complexity $O(n)$.

[4] Marr, actually, distringuished between three levels of description; the computational, the algorithmic and the 'hardware' implementation. However, the 'hardware' implementation is concerned with the biological realisation and need not trouble us.

and work its way down to the physical realisation as this can give us a better understanding of the problem at hand. The premise of this position is that there can be multiple algorithmic implementations all of which can solve the same problem but the computational problem remains invariable.

An ad-hoc sketch of the proposed *computational level* description of the memorisation problem works in two stages. First, the brain extracts small sequences of letters (i.e., the chunks). This involves things such as detecting how frequent the chunks are, how long, how complex and storing them in a temporary buffer. The second stage involves operations applied on the output of the first stage either by combining the stored chunks into larger ones or discarding them from memory. This level of explanation is void of any algorithmic details which govern, say, the space of the temporary buffer or the nature of the operations which take place during the second stage. However, knowing more or less what are the components we can implement many solutions which make different assumptions and lead to potentially different results.

We illustrate the connection between the *computational* and the *algorithmic* level by developing two models which encapsulate the computational desiderata of the proposal above, achieving similar results but making different predictions as to what would be learnt *implicitly*. Firstly, we adopt the view proposed by Perruchet & Pacteau (1990) and Servan-Schreiber & Anderson (1990) that during the familiarisation phase participants *chunk* the input strings, that is, they decompose them to smaller units which can be stored more efficiently in some temporary buffer (e.g., the working memory).

The first model we consider assumes that participants keep track of the frequencies of increasingly large chunks (uni-, bi- and tri-grams) from every training pattern they encounter. For example, upon seeing the sequence TPTS the following subsequences are stored: 1: T (twice), P, S; 2: <s>T, TP, PT, TS, S</s> and similarly for the trigram case (<s> and </s> are the start and end of the string markers, respectively). This frequency counting is supposed to be unconscious (Ellis, 2002, p. 146). During testing, participants can estimate the probability of a novel instance as the product of the probabilities that make up that specific sequence (1.1) as follows:

$$P(s) = \prod_{n\text{-gram} \in s} p(n\text{-gram}) \tag{1.1}$$

where the probability of the sequence is the product of the probabilities of the individual n-grams. This computational model does not make any assumptions about the implementation as, for example, how many chunks can the buffer retain or whether simpler chunks (e.g., TT) will be easier to retain than more complex (e.g., XVPT). Similarly, it does not make any

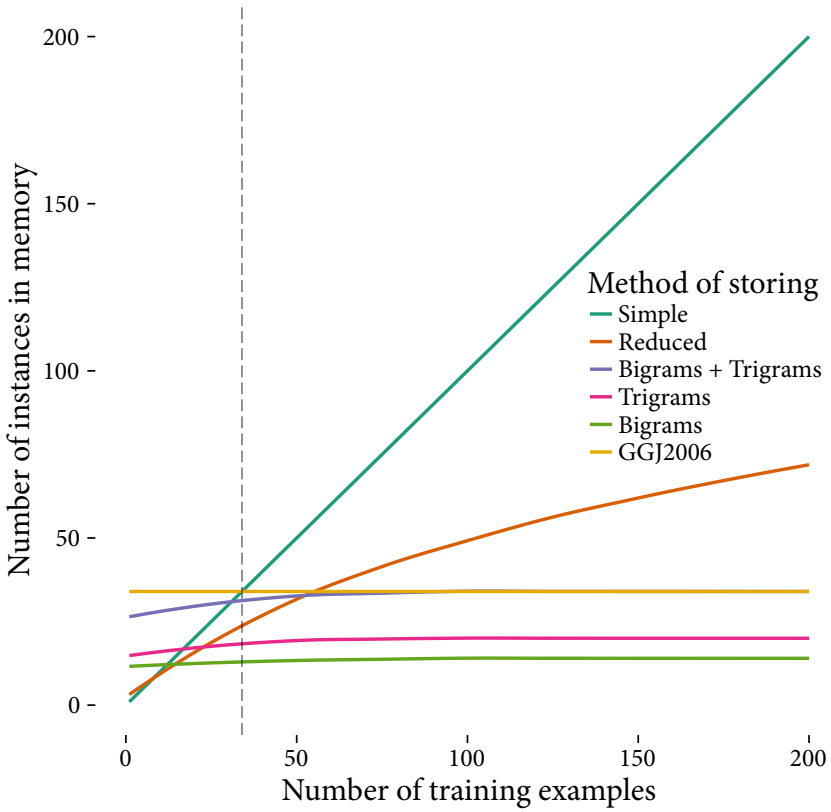


Figure 1.2 Illustration of the memory cost for different methods. 'Simple' refers to remembering every instance from the training set. 'Reduced' refers to storing a simplified version of the training examples (e.g., 'TVXXXXS' → 'TVXXS'). Bigrams, Trigrams and Bigrams + Trigrams chunk the input strings tuples of two or three characters (i.e., 'TVXXS' → 'TV', 'VX', 'XX', 'XS') or a combination of the two. GGJ2006 refers to the Bayesian model introduced in Goldwater et al. (2006) (see text for explanation). Because this model 'stores' the entire sequence in memory where the chunking occurs, it is not possible to see the size of its vocabulary as more words are introduced (see text for explanation).

assumptions as to whether it is possible for some chunks to be forgotten or misremembered. A similar n-gram based model proposed by Perruchet et al. (2002) called PARSER makes such assumptions about the cognitive implementation.

Fig. 1.2 shows the number of different chunks abstracted from the above grammar as a function on training examples encountered. Evidently, *chunking* alleviates the computational overhaul from remembering specific instances, which grows as we get more examples, to the complexity of the grammar, which grows with the number of transitions in the FSG. The space cost is, therefore, reduced from $\Theta(n)$ to $O(\log n)$.

An alternative chunking algorithm proposed by Goldwater et al. (2006) (henceforth GGJ2006) (also in Goldwater & Johnson, 2005) has achieved excellent results in modelling word segmentation in children (Frank, Goldwater, Mansinghka, Griffiths & Tenenbaum, 2007; Goldwater, Griffiths & Johnson, 2009). This chunking model attempts to find the best balance between *fit to the data* and *parsimony*. *Fit to the data* in the present context would be remembering each string encountered during the training phase. We have already seen the benefits and limitations of such a model. *Model parsimony*, on the other hand, is taken to mean, that one should choose the simplest possible model; in this context, this could be a model that accepts any string containing the letters used during training. While this performs considerably worse in the recognition task, it comes at a space cost $\Theta(|\mathcal{V}|)$ (where $|\mathcal{V}|$ is the size of the vocabulary).

The GGJ2006 model starts by storing all the training sequences in memory placing boundaries at random positions. This way both *fit to data* and *parsimony* are quite low. At every iteration, the model makes tiny adjustments (either by removing one boundary or by introducing another) and then assesses whether this improved fit or parsimony. Evidently, the model does not pose any restrictions on string length. At the end of the iteration, the model will end up with *just enough* chunks that would enable it to predict any novel sequence. During testing, we can assess this model comparing the stored chunks to the novel strings.

There are many ways to evaluate the success of these models; one would be to see if they make similar predictions regarding whether grammatical strings would be preferred more than ungrammatical. To this end, we generate 240 strings (trained on 40, generalised on 200) of length between 5 and 12 characters using the grammar in Fig. 1.1.[5] For the n-gram model, we retain only bi- and tri-grams so as to achieve comparably sized chunk inventories with the GGJ2006 model. We can then compare the probabilities for the grammatical versus the ungrammatical sets. Indeed, both models predict that grammatical strings would be more probable than ungrammatical ones. The differences between the probabilities of the grammatical strings and their ungrammatical counterparts were highly significant† in both cases ($t(306.33) = 18.208$, $p < 0.001$ for the n-gram model and $t(238.63) = 51.426$, $p < 0.001$ for GGJ2006) (although the effects were higher for the n-gram model as revealed by a `model` $\times$ `grammaticality` interaction $F(1, 796) = 29.92$, $p < 0.001$).[6]

[5] In the original 1967 experiment, Reber used 34 strings of length between 6 and 8 characters, but increasing the number of examples and the size of each example illustrates the better the memory advantages. Furthermore, subsequent experiments used a setup similar to the one we adopt here (Nieuwenhuis et al., 2013, for a more recent example).

[6] In order not to disrupt the discussion, we reserve the † (dagger) symbol to indicate either some statistical analysis or generation of artificial data without providing complete details. Fuller descriptions of their settings can be found in Appendix A.

Table 1.1 Most probable chunks for each of the two models. The GGJ2006 model by not imposing length constrains to the n-grams adds chunks to the inventory which are longer than three characters.

GGJ2006	TRIGRAM
xxvps	vxx
xvpxvs	<s>vx
tppt	ptx
xvps	vvp
xvpxxvs	ttx

<s> denotes start of string.

The results presented above show that the two computational models solve the same high-level problem rather efficiently. The comparison shown in Fig. 1.2 illustrates that by chunking one needs to retain only a fraction of the strings to be able to both recall grammatical strings and generalise to novel patterns. The algorithmic details, however, were markedly different between the two models. While for some this might be an irrelevant issue (cf. Marr, 1982) as we *should* be interested in solving the high-level computational problem, for the implicit learning research this can be quite problematic. Table 1.1 shows the most probable internal representations formed by each of the two models; apart from the differences in string length caused be their internal specifications, they are also markedly different in what they consider would be predictive in a novel sequence. The GGJ2006 model tries to find a configuration of sequences such that as many chunks will have a high chance of re-appearing on any novel string, whereas the n-gram model merely distributes the probability mass around a few high probability transitions (e.g., x → x) assigning small probabilities to the rest of the sequences. Table A.1 shows the transitional probabilities of the FSG in Fig. 1.1. Note that although the transitional probabilities are almost uniform, the data displayed to the participants during the experiments were not balanced (at least, in the initial experiments Reber, 1967, but see Knowlton & Squire, 1994) so from the perspective of the participant different n-grams might appear as more probable. Further to that, as Perruchet & Vinter (1998) show, performance issues relating to memory encoding might bias participants away from certain n-grams.

Before moving on the examination of what kind of computational model would be more beneficial for the implicit learning researcher let us examine what the above comparison added to our understanding of implicit learning. In other words, why bother with the computational level of description *at all*? After all, none of the two models seems to be an accurate depiction of what is happening in the mind during these tasks. Both models view the whole process as

an efficient way to *memorise* incoming information. A consequence of this is that some of the resulting strings (i.e., the chunks) are abstracted away and re-used to judge the grammaticality of novel strings. What we view, therefore, as implicit learning are the traces of this process. Researchers in the field (e.g., Cleeremans, 2014 and Cleeremans, 1997) espouse this view that implicit learning is the by-product of *information processing* seeing the mind as always learning from incoming information. In §1.3 we explore this idea further laying the computational framework within which the above can be realised.

Despite the valuable insight the computational level gives us, the implicit learning community is more interested in the *internal representations* these models form during learning. The reason for the above is that if we are interested in exploiting IL in *any* real-world scenario as in language or skill learning, we should know the contents of the chunking inventory. The contents of interest, however, are given to us by a model which makes concrete algorithmic assumptions contrary to the models described above. Consider the following example for a moment; the above streams are sentences from an unknown language we are trying to learn. Under the n-gram model, we would be able to distinguish seemingly ungrammatical sentences, but we would not be able to detect any words. Within the GGJ2006 model, on the other hand, we have a –statistical– notion of 'word' as a sequence of symbols that re-appears at different points within a corpus.

1.3 The computational framework

Consider that we perform a series of artificial grammar learning experiments such as the ones shown above gathering behavioural data using different manipulations. For example, we might manipulate the *average string length* and explore how this affects generalisation accuracy (as in Miller, 1958). Alternatively, we can construct grammars of different complexity (by introducing more nodes and vertices to the grammar) to see how this affects learnability (Mathews et al., 1988); lastly, we might also change the frequency of particular n-grams in the stimuli and examine how sensitive learners are in subsequent generalisation tests (Knowlton & Squire, 1994). The results obtained can be relevant to the research community as they can be used to build *theories* of *how* or *when* implicit learning would arise. Take the last manipulation in the context of the computational models sketched above; knowing that participants are sensitive to n-gram frequencies tells us that they are not necessarily learning the transitions from one node to the next but that they are chunking their input into smaller units. The probability that a participant will remember a particular chunk will then be a function of that chunk's frequency during training. Building a computational model such as the ones shown

above around this idea helps us go beyond the experimental data and explore what enables the computational model to generalise.

The limitation of doing things this way is apparent when we start asking questions relating the behavioural results to the rest of cognition. What *attentional* or *working memory* resources do we need to carry out the tasks? How does this relate to *long-term memory*? What is the role of individual differences? The inability to readily give answers to these questions comes from probing a single facet of human behaviour leaving many potentially critical parameters to chance (Sun, 2008). We can face this *de-contextualisation* by relating our results to an underlying computational framework which provides a canvas for linking the behavioural results to the rest of cognition. *Computational frameworks* are theoretical constructs which do not readily make predictions for specific tasks (Anderson, 1990) but function as *primitives* which can be elaborated to make concrete predictions about specific experiments.

To illustrate this point, consider the *Bayesian framework* in psychology (Anderson, 1990; Tenenbaum, 1999; Tenenbaum & Griffiths, 2001). According to proponents of this framework, the mind is constantly 'bombarded' by data $\mathcal{D}$ and tries to uncover the underlying mechanism responsible for producing them. The mind does so by picking out the best hypothesis h from the countably infinite set of all possible hypotheses $\mathcal{H}$. The hypothesis chosen should maximise the *fit to the data*, that is, the probability of the data given the hypothesis $p(\mathcal{D}|h)$. However, some hypotheses might be overly intricate, and while fitting the data better, they are intuitively less probable. Each hypothesis is then modulated by a quantity called the *prior* probability $p(h)$, which ensures model *parsimony*, to determine the best hypothesis that generated the data. In other words, world experience (i.e., our intuition about the probability of each hypothesis) shapes[7] what we learn and what to expect and is enrichened every time by incoming data. The above can provide a canvas for the researcher which can be adapted to different tasks. For example, in GGJ2006 $p(\mathcal{D}|h)$ can be some form of entropy which we need to minimise while $p(h)$ can penalise longer strings. What is common to all Bayesian models of cognition is that the hypothesis chosen will depend on its likelihood (i.e., the fit to the data) times its prior probability.

Choosing amongst the available computational frameworks to model implicit learning data while crucial is not a straightforward issue. Most models which aspire to provide an explanation of psychological phenomena take Marr's division of levels of description or J. R. Anderson's *ideal observer* as their starting point targeting the computational level of description (Anderson, 1990; Frank & Tenenbaum, 2011; Griffiths, Steyvers & Tenenbaum, 2007; Tenenbaum, Perfors & Regier, 2011a; Xu & Tenenbaum, 2007) solely. As we saw above,

[7]This circularity in reasoning has stirred criticism against the Bayesian framework (e.g., Bowers & Davis, 2012), see Griffiths, Chater, Norris & Pouget (2012) and Hohwy (2013) for theoretical justifications.

this way they carefully describe the goals of the system, provide a better insight into the problem, lacking, however, rigorous algorithmic descriptions, important in implicit learning.

Algorithmic models are abundant in psychology since they are interested in matters of *efficiency*, *degradation of performance*, *time it takes to solve a problem* (Rumelhart & McClelland, 1985) but look more at very specific phenomena (e.g., John Morton's *logogen* theory of word recognition, Morton, 1969). A good compromise between the two desiderata, that is, a general framework which makes concrete algorithmic implementations, are the *parallel distributed processing models* (Rumelhart, McClelland & PDP Research Group, 1986b) or the *Naïve discriminative learner* (Baayen, 2010). These frameworks bring together domain-general learning mechanisms with concrete assumptions about the representations they are using and how these representations are transformed to carry out the computational task. The framework chosen throughout the present thesis extends the ideas of the Parallel Distributed Processing (PDP) models introduced in the 1980s and collectively known as *connectionism*. In what follows, we will be introducing some aspects of the original theory as well as where our extensions lie using state-of-the-art machine learning methodologies. Subsequently, following the ideas of Cleeremans (2014), we will be arguing why this approach is the most appropriate in the present context and how implicit learning phenomena are bound to arise in the current context naturally.

Parallel distributed processing models, also collectively known as *connectionism* is a psychological framework within which processing and learning occur through a bundle of simple and interconnected units. We identify the units as *simple* because the only thing they are concerned with is to modify parts of their input, either by inhibition or amplification, passing it to other units to perform some sort of computation. The second point about the units being interconnected is as important as one unit by itself has quite limited learning capabilities; many units, however, can discover regularities in the input enabling the system to make predictions from 'noisy' input. Parallel distributed processing models, or artificial neural networks or connectionist networks (all the terms have come to be synonymous) are general-purpose learning mechanisms inspired after the biological networks of neurons in the brain. Their general purpose mechanisms render them suitable for modelling a number of different psychological phenomena with some adaptations. The purpose of the following attempt of a sketch of the connectionist framework is twofold; (1) to explain concepts introduced later on in §2.3 and Chapter 5 and (2) to explain *why* implicit learning naturally arises in this framework. However, for a complete examination, one can consult Rumelhart et al. (1986b) for some early results, Elman, Bates, Johnson, Karmiloff-Smith, Parisi & Plunkett (1996) for a more developed view of connectionism or Goodfellow, Bengio & Courville (2016) for some recent developments in artificial neural networks.

From an engineering point of view, artificial neural networks are complex statistical models performing, but not limited to, linear or logistic regression. That is, given an input set of predictor variables such as an array of colour values representing an image and a corresponding output such as the name of the animal shown in the image, neural networks, using associative learning mechanisms, learn the statistical constraints present in the input material, enabling them to generalise to novel input. Neural networks are, thus, *function approximators*. Concretely, let $f^\star$ be an unknown function which maps some multi-dimensional input $\mathbf{x} = x_1, x_2, \ldots, x_N$[8] (e.g., the image) to an output y (e.g., the name of the animal).[9] Let also y_t be the target label from the space of all possible labels Υ. Finally, to keep the neural network analogy, we call the slots for the input or output patterns *units* and their connections *weights*. The underlying function $f^\star$, therefore, given input $\mathbf{x}$ predicts the target output y_t. The goal of the network, on the other hand, is by 'looking' at input-output pairs to be able to predict what the output of the original function would be. That is, they learn a function f such that $f(\mathbf{x}) \approx f^\star(\mathbf{x}) = y_t$.

The way neural networks learn to make this prediction is by associating parts of the multi-dimensional input representation to the output. The network keeps track of how small variances in the input affect the output finding ensembles of units which are more predictive of certain outputs. In order for the network to do this association, it has to learn some *parameters*, the weights of the neurons. In short, consider we have to do a classification with two input units (e.g., height and weight) and one output (e.g., gender). These weights describe how much each predictor is associated with the output (e.g., high values for height would mean that height best determines gender). If we wanted to make a prediction for more output variables, then we would need one weight from each unit of the input to each unit of the output. From now on, we will use $\vec{\theta}$ to denote collectively those parameters the network needs to learn.

Approximating $f^\star(\mathbf{x})$ with $f(\mathbf{x}, \vec{\theta})$ means that we have to somehow *learn* those parameters $\vec{\theta}$. The networks learn those parameters by constantly making predictions given an input and then make tiny adjustments until they predict the correct output. To do so, the network needs to have an internal metric of 'how well it is performing'. We call this metric, the *cost function*. If, for example, given the image of a 'cat', the model predicts 'dog' it needs to know about the mistake so that the next time it predicts the correct label. The *cost function* is then a function which compares the predicted output from the model to the given 'golden' output and returns an estimate denoting performance. Depending on the network's task (e.g.,

[8] We use linear algebra notation instead of the most common –in psychology– sum notation. We refer the readers who are not used in this notation to Jordan (1986) for an excellent introduction. We also provide short explanations on the notation at the end of the thesis.

[9] For simplicity we consider the case where the network predicts a single output. However, neural networks are perfectly capable of learning multivariate functions of the form $f : \mathbb{R}^N \to \mathbb{R}^M$, $M \geq 1$.

classification, regression, sequence prediction) different cost functions might be appropriate. In §5.3.1 we discuss the workings of a couple of cost functions, and why they are appropriate given the tasks, we attempt to model.

How does learning take place in a neural network? Since we have access to positive examples (i.e., input-output pairs) as well as a metric of how well our system is performing we need an algorithm so that at each step $\vec{\theta}$ moves to a direction that minimises the cost function. Networks with simple architectures can make use of very simple learning algorithms such as adding or subtracting a constant value to each weight (Rosenblatt, 1958). On the other hand, more complex networks keep track of how *each* weight in $\vec{\theta}$ affects the error function making the appropriate adjustments. Having computed this quantity the algorithm makes tiny updates to its parameters such that at the next iteration the error is reduced. Computing how each weight affects the cost function has stirred criticism against artificial neural networks. The reason for this is that this implies we compute the partial derivative of the cost function w.r.t each weight in the network, $\frac{\partial E}{\partial \vec{\theta}}$. Critics argue that this computation cannot be implemented by the brain's cortical structures. Hence artificial neural networks do not provide an accurate description of the brain's learning mechanisms. We do not consider this to be a problem in the present context as according to Marr's division of levels we are looking at the *algorithmic* level of description and not the *biological*. However, one can consult Rumelhart & McClelland (1985) and Hinton (2014) for arguments linking the two levels as well as Scellier & Bengio (2017) for a potential biological realisation.

Concretely, the networks we have introduced so far compute the following function; $f(\mathbf{x}) = \mathbf{W} \cdot \mathbf{x} + \mathbf{b}$. For now, we will not concern ourselves with the bias term $\mathbf{b}$ as this is used only to shift the output of the multiplication. $\mathbf{W}$ is a matrix of learnable parameters (neurons) associating the input to the output $\mathbf{W} \subset \vec{\theta} \in \mathbb{R}^{D_0 \times D_1}$ (where D_0 and D_1 are the input-output dimensions, respectively). Apart from various regularisation terms we can introduce, learning a matrix $\mathbf{W}$ such that $f(\mathbf{x}) \approx f^{\star}(\mathbf{x})$ is doing simple regression. What distinguishes neural networks from simple regression is that they can increase their *representational capacity* by introducing intermediate outputs before their final prediction. In short, in the case where the input is either 'too noisy' or does not contain enough information the data might not be *linearly separable* making prediction harder. In that case, the network *learns* an intermediate *implicit* representation which is a transformed 'version' of the input passed through a non-linear function (called the *activation function*). The network function we end up with is $f(\mathbf{x}) = \mathbf{W}_{ho} \cdot \mathbf{h} + \mathbf{b}_h$, where $\mathbf{h}$ is the hidden layer computed as $\mathbf{h}(\mathbf{x}) = \mathbf{W}_{ih} \cdot \mathbf{x} + \mathbf{b}_h$ and σ is a pointwise non-linear function (e.g., tanh). This way neural networks are able to model functions of arbitrary complexity. Geometrically, intermediate –hidden– layers are transformations of the input (or of the layer before then in the case of multi-layer neural networks) which alter the

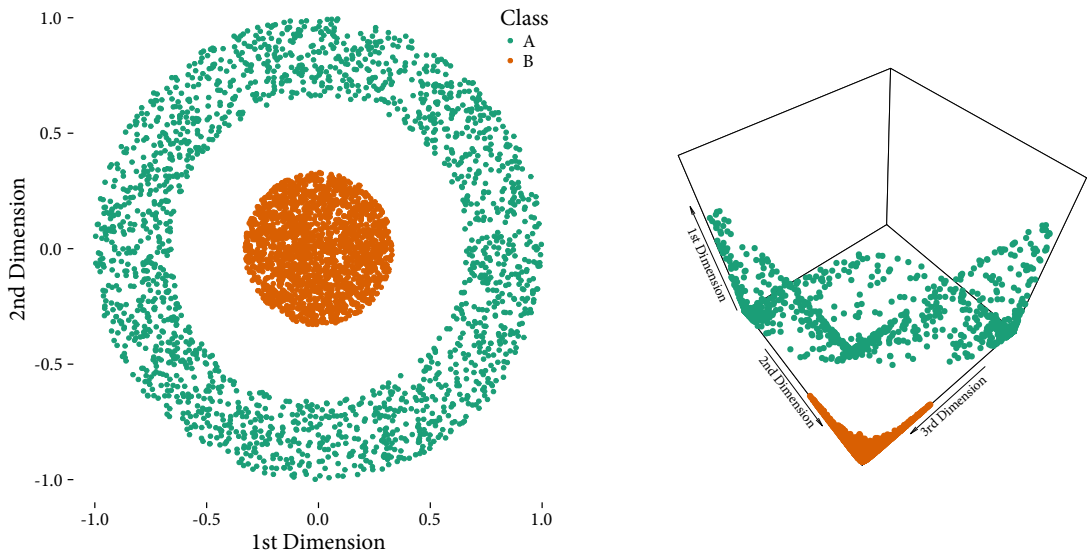
dimensionality of the input and possibly performing an element-wise non-linear operation (such as a sigmoid or a tanh function). These operations either stretch or squish the input space such that the function becomes easier to compute.

To demonstrate the benefit of this operation consider the toy dataset† in Fig. 1.3. Fig. 1.3a shows points randomly dispersed in a two-dimensional plane. Points in the outer ring belong to class A and points in the circle in the centre in class B. Elementary geometry shows that the task is not simple for a linear classifier as the cutoff margin is circular. No matter how much we stretch or squish this 2D space, there is no way to separate the two classes.[10] However, introducing a non-linear hidden layer the problem becomes trivial (Fig. 1.3b) as there clearly exists a plane which easily separates the data. This demonstrates a critical point for later; while the raw input might contain all the information we need to carry out the task, the underlying system might need to apply some internal processing to simplify its job. This operation increases the number of parameters $\vec{\theta}$ in the model as we now need coefficients for both the raw and the transformed input.

To take a linguistically motivated example of an unknown function $f^{\star}$ that can be approximated by such a system we take the Past Tense formation in English (Rumelhart & McClelland, 1986). Having only access to the phonological representation of the verbs, as, for instance, /ˈpleɪ/ (play), we need to learn a function F such that F : Present Form → Past Form. This function takes as input a low-level (surface) phonological representations of the verbs in the present tense as above and it learns to associate this with a similar representation in the output layer (Table 6.1 offers an example of such phonological representations). Given appropriate representations, the problem is reduced to multivariate linear regression were the network given a vector of real values for input has to predict another in the output. The extent to which it can achieve this depends on (1) the scheme we employ in capturing phonology in the input as well as (2) how recoverable the rule is only from phonology. Rumelhart & McClelland (1986) and Joanisse & Seidenberg (1999) show that both regular and irregular past tense formation can be induced solely by the phonology of the verb (but also see, Tyler, Stamatakis, Jones, Bright, Acres & Marslen-Wilson, 2004).

This introduction has highlighted that neural networks provide a general-purpose learning framework upon which we can study psychological phenomena. The exact way we can adapt them to specific tasks depends on our specification of the abstract computational problem. Consider that participants are introduced to three new words; *dax*, *pax* and *lex* and that different known words fall in one of three categories depending on the letter they start with. This is a trivial problem of *categorisation* and would require a feed-forward network with three units in the output layer (the depth of the network depends on the input representations).

[10] See http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/ for proof.



(a) (b)

Figure 1.3 Example of an unlearnable input and its solution. (a) The class A which forms an annulus in the out and class B which is the circle in the middle. (b) The input pattern as transformed by the three-dimensional hidden layer. Projecting the input in a three-dimensional space and then applying a non-linear operation helps the network find a solution where it is easy to find a plane to separate the classes.

Similarly, if the task were to predict a scalar value (i.e., linear regression), we would only need to change the activation function in the output (and perhaps the cost function). We can also treat sequence learning as either classification or regression with the added constraint that the internal state at the previous timestep is carried over.

Even from this brief introduction, we see that there are many different ways to initialise these systems[11] by defining different topological characteristics or objectives (i.e., what they will be trying to learn). Hornik, Stinchcombe & White (1989) has famously proved that neural networks with only a single hidden layer are *universal approximators* (i.e., they can model any computable function, also see, Graves, Wayne & Danihelka, 2014). The choices we make regarding the above depend on the computational task the system is trying to solve, as well as on the nature of the data.

Researchers within the connectionist community have identified several issues over the decades which have limited computational models to toy simulations and proofs of concept

[11]See `http://www.asimovinstitute.org/wp-content/uploads/2016/09/neuralnetworks.png` for a 'mostly' complete chart of the different neural architectures.

rather than realistic datasets. One major early issue was *catastrophic interference* (Ratcliff, 1990) where a network in order to learn *novel* information *overwrites* previously learnt ones. This and other issues related to *parameter optimisation* (i.e., techniques which are responsible for learning the values of the parameters) have cast doubt on neural networks' ability to provide a scalable account of human learning mechanisms. Recent methodological progress, however, has renewed the interest not only for engineering purposes (Graves, Mohamed & Hinton, 2013; LeCun, Bengio & Hinton, 2015) but also for cognitive scientists (Hinton, 2014). Duchi, Hazan & Singer (2011) and Zeiler (2012) have proposed adaptive learning of weights, and Glorot & Bengio (2010); Hinton & Salakhutdinov (2006) use different *initialisation* techniques which provide sophisticated algorithms to initialise the weights randomly or by iteratively pre-training the different layers can sidestep the issue of getting stuck in local minima. These advances together with more sophisticated architectures (Chung, Gulcehre, Cho & Bengio, 2014; Hochreiter & Schmidhuber, 1997) have led to an explosion of theoretical and practical applications of connectionist networks under the umbrella term of *deep learning* (LeCun et al., 2015) sidestepping many of the early shortcomings of neural networks. The present thesis while adopting the original ideas provided by *parallel distributed processing models* also adopts the methodologies used in natural language processing, image and speech recognition to provide a better view on implicit learning.

What advantages does this computational framework offer us in exploring implicit learning? Firstly, contrary to classical models of cognition (Fodor, 1975; Fodor & Pylyshyn, 1988; Newell, 1990), in connectionist networks learning is not a specialised operation. The dominant metaphor in 'classical' models of cognition, is that humans are equipped with a database of rules or facts and that any incoming knowledge either adds elements to the database or combines these elements to generate more complex representations. This way learning is an accumulation of information gathered from the environment carefully compiled by a set of possible operations defined by human cognition (see, e.g., Anderson, 1990). In §5.8.3 we explore the idea of how *semantic implicit learning* could arise in these models. Cleeremans (2014) has shown that these assumptions of classical computational frameworks of cognition cannot explain implicit learning processes (see also Bates & Elman, 1992 for a more detailed comparison between classical models of cognition and early connectionist systems). The problem as he identifies is that knowledge in such systems needs to be represented symbolically and subject to conscious experience. In other words, knowledge in classical approaches is knowledge 'we know we have'. In artificial neural networks, on the other hand, where learning is largely a by-product of information processing knowledge is causally efficacious yet unavailable to conscious experience (see also Clark & Karmiloff-Smith, 1993).

A second attractive quality of artificial neural networks in exploring implicit learning is their inability to express their internal states. We have already seen that a core component of implicit knowledge is that it influences processing without being verbalisable. Participants can carry out the artificial grammar learning tasks, without, however, being able to explain the rules of the grammar. Artificial neural networks tie quite well with this idea in that knowledge does not form an *object of representation* by itself. In other words, contrary to symbolic systems, connectionist networks do not encode knowledge in the form of *rules*. Instead, knowledge exists in the configuration of weights the network has figured to carry out the task.

The complete lack of concrete rules has been considered problematic by critics (Fodor & Pylyshyn, 1988) as, undoubtedly, humans can exploit known regularities to perform actions. Lack of rules, however, is not tantamount to lack of rule-like behaviour. That is, although we do not encode any symbolic rules in these systems, this does not mean that they cannot express rule-like behaviour. §A.2 presents the generative process responsible for placing the points in Fig. 1.3a. While the neural network does not have knowledge of these rules, it can perfectly classify the points in a rule-like manner.

1.3.1 Representations

Let us take a step back and examine what exactly is the input to the above systems. Keeping in mind the idea that learning is a by-product of information processing, the primary input to these models should be a description of an entity we come in contact with and is found in the environment as, for example, sound or light. This 'raw' input description will contain, however, an immense amount of information, some of which is going to be predictive for the task while the rest will be environmental 'noise'. The models introduced above can cope with that noise to a large extent and transform the input into an internal representation useful to carry out the task. There are two questions we explore in the present section; (1) how can we distill environmental information into a format which would be usable for the system to carry out a specific computation and (2) how can we describe entities at different levels of abstraction (e.g., animal as opposed to dog)?

Take the human visual system for example. Despite the complexity of the operations involved in detecting edges, hues, shapes and orientations, the initial 'raw' representation which is given from the physical world is quite straightforward. Briefly, the photoreceptors in the retina detect light intensity values.[12] Knowing, therefore, the light intensity value at

[12]We make the simplifying assumption at this point that humans can see only one colour (e.g., in grayscale). This assumption is taken not to clutter the text with more complicated notation. In any way, the above logic trivially extends to multicolour recognition if we accept that there are different kinds of photoreceptors sensitive to different colours.

each point in the visual field, we can construct a 'map' the coordinates of which show us the intensity values. For convenience, let the matrix $\mathbf{M} \in \mathbb{R}^{X \times Y}$ be that map where X and Y are the size of the visual field. Each element $\mathbf{M}_{ij}$ is a light intensity value ranging from 0 (i.e., white) to 1 (i.e., black). While this representation might be cluttered with a lot of irrelevant noise, the human visual system is able to extract all the relevant information in order to recognise objects, faces and so on.

In short, these *representations* which enter the system define a systematic way of describing entities or types of information (Marr, 1982) available during processing. As above, visual representations can describe light intensity; auditory representations, on the other hand, can describe the wavelengths of vibrations in the environment. There are two remarks that need to be made at this point; (1) as we noted above, environmental input is 'noisy'. In order to not clutter the model with irrelevant information which might need more time to train or provide more local minima, the input is typically pre-processed so as to reduce the amount of superfluous information contained within it. (2) the word 'systematic' is of importance in the present context. Inevitably, the description we choose is going to highlight some features of the input while push back others. For example, if we train a network to distinguish dogs from cats it might be irrelevant to include in the input that some dog breeds are more susceptible to canine epilepsy than others.

While for visual and auditory recognition processes the input representations are quite straightforward, the matter of input representation is more complicated when we look at language. Do we use 'raw' visual input (i.e., printed texts)? or should we start with speech input? do we pre-process grammatical constructions or we somehow let the system figure them out? Choosing among the alternatives implicitly subscribes us to a level of linguistic description which might or might not be appropriate in the present context. For example, *generativists* do not really care about the environmental input as this is too variable but focus more on a higher level of explanation once this input has been mapped to something more invariant (commonly called the I-Language, Chomsky, 1986). For a researcher subscribing to connectionism the primary environmental input can be very informative as seemingly complex rules such as the Past Tense formation which we saw above can be recovered solely on that level without appealing to additional mechanisms (Rumelhart & McClelland, 1986) (although see Tyler et al., 2004).

Staying on the topic of linguistic representations, the matters are complicated further when we consider *semantics*. Undoubtably, the issue of semantic representations is one of the most formidable topics in cognitive psychology. Different approaches including philosophical (Wittgenstein, 1922), psychological (Barsalou, 2008; Collins & Quillian, 1969; Rogers & McClelland, 2004) and computational (Landauer & Dumais, 1997) have sought to explore how

words relate to each other and to concepts or how, in turn, concepts relate to each other or to percepts and actions. The multimodal nature of semantic representations makes it harder for the modeller to construct a description of concepts which captures what is evoked in the brain when a target concept is seen or heard. Since semantic representations are at the heart of the present thesis, we devote a few paragraphs introducing the different ways we can represent semantics which can either emphasise the abstract conceptual structure or the associative relations between the words. This can by no means be considered a complete review of the ways we can capture semantic relations. Chapter 2 goes into detail on how to extract semantic representations by looking at word co-occurrences. Furthermore, in §3.3.2 we go into more detail on how to turn the representations below into an appropriate input for the networks.

Word Association Norms

Word Association Norms (AN) focus on describing how words are associated with each other. This gives a *shallow* semantic representation in that we do not necessarily take into account the *nature* of the relation between the words. Generally, such association norms are compiled by asking participants to produce the first word that comes to mind that is meaningfully related to a target. For instance, a participant might encounter,

GRADUATE .....

to which they might respond, *student*, *school* or *degree*, in which case these would be considered associates to the target GRADUATE. The semantic representation can then be constructed by looking at which words relate to the target. In this case, the input to the model can be a $|\mathcal{V}|$-dimensional binary vector (where $|\mathcal{V}|$ is the number of words used as targets) where the non-zero values indicate that a word is associated with the target. To get an even more accurate image of the relationship the target word has to its associates we might also want to weigh the potential responses according to how many participants gave that answer. For example, if 24 people responded *student* in the above example, out of 148 this gives a weight of .162. The corresponding element, therefore, in the vector representation is going to be .162 instead of 1 indicating a weak relationship between the two words.

These association norms have been an indisposable tool for cognitive psychologists because of their coverage and the ease with which they can be generated. The University of South Florida Free Association Norms (Nelson, McEvoy & Schreiber, 2004) we use later on contain targets for 5019 cue words and are commonly used in designing stimuli lists (e.g., Hutchison, Balota, Neely, Cortese, Cohen-Shikora, Tse, Yap, Bengson, Niemeyer & Buchanan, 2013) or as a benchmark for evaluating systems which automatically generate such representations (Kiela, Hill & Clark, 2015b). More recently, Deyne, Navarro, Perfors & Storms (2016); Deyne,

Navarro & Storms (2012); Deyne & Storms (2008) have extended this line of work by providing a extensive database of association norms for both English (Deyne et al., 2012) and Dutch (Deyne & Storms, 2008) which lead to better predictions of lexical access and semantic relatedness, particularly for words which are weakly related.

Semantic Feature Norms

Semantic feature norms focus mainly on the relations between concepts and percepts and actions or to other concepts. Examples of such relations are that cats have tails (concept to concept) or that cats are independent. These representations do not worry about the relations between *words* and the *concepts* they refer to. This *indirect* relationship to language enables them to go beyond the mere associations captured by the word ANS. Similar to word ANS, *semantic feature norms* are collected by asking participants to list properties for a target word. Participants are instructed to include properties such as: physical, how the concept referred to by the target words looks, sounds, smells, feels or tastes.

Semantic features are commonly used or assumed to exist in several theories of categorisation and conceptual representation. For example, in exemplar theories of categorisation (Nosofsky, 1986), participants are assumed to attend to correlations of features, and how these are predictive of the category a concept falls in. In formal modelling, MINERVA 2, a model of associative memory, assumes that memory is composed of empty slots which are filled with the features of the incoming probe. Incoming stimuli containing the relevant features strengthen the association of this element to the category.

In cognitive modelling *semantic feature norms* are used to model a variety of psychological phenomena. Cree, McRae & McNorgan (1999) used the semantic feature norms compiled by McRae, de Sa & Seidenberg (1997) (an earlier version of McRae et al., 2005) as input to an attractor neural network, a special type of the networks introduced above where the training pattern continues to activate the output for a few timesteps until it settles to a stable pattern in the output, to successfully model semantic priming effects. Moreover, Mirman & Magnuson (2008) looking at the effect of semantic neighbourhood density on word recognition, also used the McRae norms as input to an attractor neural network measuring the time it took for the model to settle to a pattern in the output layer. Finally, Rabovsky & McRae (2014) modelled seven N400 Event-related potential (ERP) component effects reported in the literature using, again, the McRae norms. Other commonly used feature norms include the ones gathered by Vinson & Vigliocco (2008) which described objects and scenes as well as those collected by Devereux, Tyler, Geertzen & Randall (2013).

WordNet

Similar to semantic feature norms another commonly used semantic description capturing concept to concept relations is WordNet (Fellbaum, 1998). WordNet is a large database where concepts are represented in terms of abstract propositions (to a great extent IS-A relations), as, for example, dog IS-A carnivore (see Fig. 1.4). This organisational scheme in WordNet captures the hierarchical nature of semantic relations as evidenced by developmental (Keil, 1979), reaction time (Collins & Quillian, 1969) and brain damaging data (Warrington, 1975). Again, as above, language is only indirectly addressed as all the contained words are normalised to their corresponding concepts, but contrary to the above, WordNet is hand-coded. In this way, WordNet looks more like a machine-readable dictionary/encyclopaedia than a model of semantic memory.

Because of its coverage and granularity which extends beyond what is commonly captured by semantic feature norms, WordNet is a commonly used tool both in cognitive modelling (Miller & Fellbaum, 1992) and *natural language processing* tasks (Harabagiu, 1998). Budanitsky & Hirst (2006) use WordNet and various semantic similarity metrics to evaluate how close WordNet representations fit human similarity judgements. Ó Séaghdha (2007) achieved state-of-the-art results on a compound noun learning task (e.g., *steel knife*) using WordNet representations. In Chapter 3 we examine whether WordNet representations are *also* suitable for modelling implicit learning tasks.

We recognise the importance and appropriateness of all the different paradigms to study the organisation of the semantic memory. Undoubtedly, they have different strengths, and they are likely to be more appropriate considering various tasks. This appropriateness stems from the fact that the representation we choose to use is bound to highlight some aspects of the input pushing others in the background. Semantic feature norms and WordNet focus on the relations that *concepts* establish with other *concepts* whereas word association norms remain on the *word-word* level. Furthermore, the level of *granularity* can potentially be another issue; the scope is much more constrained in Feature Norms (FN) than in WordNet. This level of details comes, however, at a computational cost as it introduces potentially irrelevant noise.

In Chapter 2 we will be outlining a more sophisticated method to extract associative relations between *words*, which extends the scope from a few words to every other word in the English vocabulary. This can potentially be problematic for reasons similar to the ones *generativists* have chosen to look at I-language instead of E-language. Looking at language usage instead of abstracting the underlying conceptual structure might lead us to spurious problems related to individual differences (e.g., linguistic knowledge, fatigue, dialect spoken,

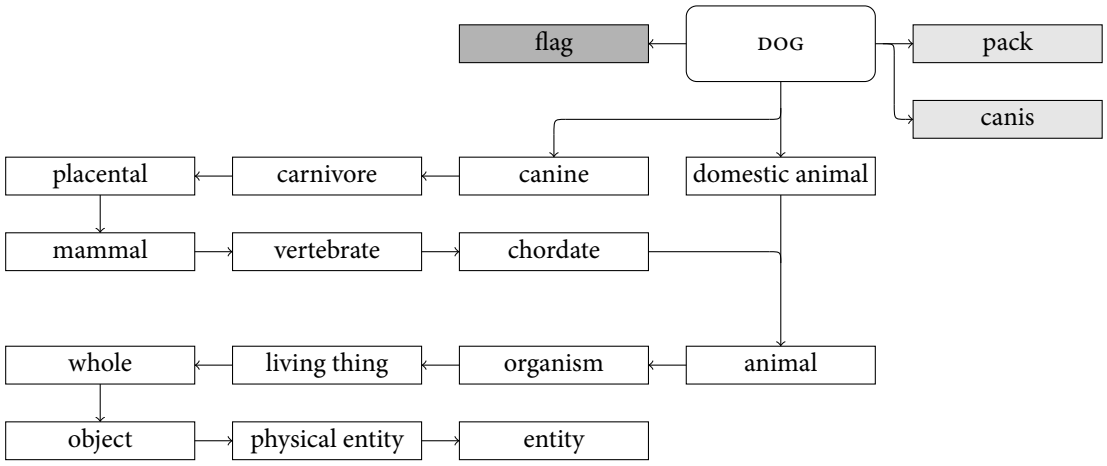


Figure 1.4 WordNet hierarchy for the lemma *dog* (synset: `dog.n.01`) showing three kinds of relations; (a) IS-A relations in white (e.g., dog is a domestic animal), (b) HAS-A relations in dark gray (e.g., dog has a tail –called a 'flag' on some breeds), (c) IS-MEMBER-OF relations in light gray (e.g., dog is member of a pack). Also, we make two remarks regarding this figure; (a) there can be synsets where there is not necessarily one path from the synset to the root (always `entity.n.01`). In cases where multiple paths co-exist, we follow them to the root filtering the duplicates. (b) We also note that for readability reasons we omit the specific sense which might cause confusion with the infrequent use of *flag* as another word for 'dog tail'. In constructing the feature matrices, however, the entire synset name was used (i.e., `flag.n.07`).

etc.). However, the models we present later on can cope with such differences and be used as a proxy to gain insight on the structure of semantic memory. Further to that, many theories of sentence and discourse understanding (Ericsson & Kintsch, 1995; Kintsch & Mangalath, 2011) identify this level of description as an important one in the early stages of language processing.

1.4 Implicit Language Learning

A significant amount of research in artificial grammar learning and especially *chunk learning* has shed light on many aspects of language learning processes (Ellis, 2015). This might seem somewhat odd as the statistical regularities inherent in the training patterns are void of any phonetic, morphological or syntactic information,[13] unlike natural languages. To see how the above results can relate to language acquisition, we need to go beyond the non-

[13]The finite-state grammar used in Fig. 1.1 implies the existence of 'syntactic' rules in the presented strings. Although an FSG can provide an adequate approximation of language on many Natural Language Processing (NLP) tasks (Kaplan, 2003), it cannot be considered an accurate description of any human language (Chomsky, 1957).

sensical patterns the participants come in contact with and look at them as *sequentially related combinations of strings*. The idea that language is the sequential concatenation of strings is quite old in linguistics (de Saussure, 1916) and although debated heavily by Chomsky (1957) has had a profound influence on cognitive psychology and especially studies on speech segmentation (Aslin, Saffran & Newport, 1998; Saffran, Aslin & Newport, 1996), learning of phonotactic constraints (Dell, Reed, Adams & Meyer, 2000; Warker & Dell, 2006) or orthographic regularities (Pacton, Perruchet, Fayol & Cleeremans, 2001).

Consider, for example, the study by Durrant & Doherty (2010); if humans are endowed with a chunking mechanism such as the GGJ2006 which can span beyond the boundaries of a single word, then high-frequency collocations will be assigned a higher probability or even considered as a single chunk. Durrant & Doherty (2010) presented participants with low-frequency collocations (e.g., '*famous saying*'), moderate-frequency (e.g., '*greater concern*'), high-frequency (e.g., '*mental picture*') and psychologically associated collocations (e.g., '*card game*'). For the computational models presented above, these collocations would maximise the probability of appearance of the second word (i.e., $P(\text{game}|\text{card})$) but do not occur as much so as to be considered a single chunk (i.e., 'card_game'). Regarding the behavioural results, Durrant & Doherty (2010) indeed found that for the last two conditions there was significant priming for the second word in the collocation indicating that the participants were sensitive to the transitional probabilities from the first to the second word (see also, §5.2). The low probability cases, on the other hand, there was no priming on the second word, probably because the first word was not as predictive. In other words, if we construct a probability distribution of the words that can follow the first word in the collocation, in the low-frequency cases this would be more uniform, whereas in the high-frequency cases the probability mass would be placed on a single word.

The computational models presented above (especially the GGJ2006) make the assumption that frequent enough collocations would receive high-transitional probabilities (cf. (1.1)) but *very* frequent collocations would merge into a single chunk. In a collocational priming experiment as in Durrant & Doherty (2010) this would slow down the reaction to the second word as it would be harder to recognise it as a component part. Indeed, Kapatsinski & Radicke (2009) performed an experiment where the participants were instructed to monitor the word *up* in phrases such as *give up* or *keep up*. Priming effects were correlated with the transitional probability for frequent collocations, whereas for the *highest*-frequency collocations (e.g., set up) there was a slowdown as predicted by the above computational models. Presumably, the slowdown happens because collocations such as 'set up' have merged through frequent usage into a single chunk (i.e., 'set_up') which participants have to segment and then parse its component parts separately.

Learning of frequencies and transitional probabilities are two prime examples of implicit learning in natural languages. Acquiring these association statistics is an unconscious process which happens only through exposure to 'raw' linguistic input by gradually changing the synaptic connections in the brain. Indeed, by now we have evidence from developmental psychology that the speed of word recognition at infancy can be a strong predictor of linguistic abilities during late childhood (Marchman & Fernald, 2008). Further to that, a substantial amount of theoretical work has linked frequency and statistical based effects to language acquisition (Bod, Hay & Jannedy, 2001; Bybee & Hopper, 2001; Diessel, 2007, for comprehensive reviews).

Ellis (2005) reviewing a bulk of evidence on L2 form-meaning connections in relation to implicit and explicit learning highlights a significant interaction between the two modes of learning. Explicit learning, such as classroom instruction, can be used to guide the focus of language learners to certain form-meaning constructions. Once this happens, however, unconscious, implicit learning mechanisms constantly update the frequencies and probabilities of this mapping to facilitate subsequent processing and learning. Indeed, there is abundant literature that L2 learners also utilise frequency distributions and co-occurrence statistics in a similar manner to native speakers when processing formulaic phrases in their second language (e.g., Conklin & Schmitt, 2007; Jiang & Nekrasova, 2007).

Despite the differences between Artificial Grammar Learning (AGL) and natural language acquisition, we see that the former provides a reasonable abstraction on which we can study the micro-processes that guide the latter. Indeed, the AGL studies along with the idea that we use *chunking* as a fast and efficient way to retain in memory linguistic information have formed the basis of a recent theory put forward by Christiansen & Chater (2016a) (also in Christiansen & Chater, 2016b). According to this, learning a language means learning to efficiently process the continual deluge of linguistic input. The solution in this theory is given by the assumption that the mind constantly chunks its incoming input. Although the frameworks differ, this underlying idea that *language acquisition is nothing more than learning to process* (Christiansen & Chater, 2016a, p. 114) is inherent in the connectionist framework presented above.

1.5 Semantic Implicit Learning

The learnability of syntactic and morphophonological regularities have long been the focus of applied linguistics pushing semantics to the periphery. We can maybe trace this tendency to the persistence of theoretical linguistics to consider syntax and phonology as the primitive notions (i.e., the 'building blocks') which should underpin a linguistic theory (Chomsky,

1977). In this view, semantics is nothing more than a 'form of syntax' providing additional constraints during the underlying computation of the sentential structure. The reasons why Chomsky and a large cohort of theoretical linguists have taken this view and how this has changed over the years fall outside the scope of the present thesis (for an early review see Katz, 1980 and Partee, 2014). Whatever one's views about linguistic theory, nevertheless, it remains an open question whether semantic knowledge can influence language acquisition. There are two issues we need to resolve before we begin our survey of semantic implicit learning studies; (1) what do we mean by *semantic regularities* and (2) what kind of semantic knowledge can determine whether a particular linguistic structure is learnable or not.

Semantic regularities are those co-occurrence patterns which do not necessarily involve syntax, morphology or phonology but include some aspect of the *meaning* of the word in question. For example, the gender system in Italian is mostly determined by the phonological characteristics of the nouns (e.g., masculine nouns end in *-o* and feminine in *-a*). Furthermore, the construction that a particular determiner is placed before the noun in English is an example of a syntactic regularity. While perhaps not as overtly, semantic regularities are ubiquitous in the languages around the world. Xhosa and Bantu languages in general, for instance, distinguish 15 noun classes according to their meaning (Denny & Creider, 1976). In Xhosa, abstract nouns are prefixed by *du-* (class XI) whereas humans and mostly animate concepts are prefixed by *mụ-* (class I); other elements of the sentence that should agree with the noun are also prefixed by these morphemes. While, therefore, there is a distributional relationship between the morphological marker and the noun the co-occurrence relationship is controlled by semantics. In a similar vein, in Chinese *noun phrases* (NP) with either a quantifier or a numeral, the speaker must include a specific grammatical marker (called the classifier) between the noun and the quantifier/numeral. The choice of this classifier, however, is conditioned on the meaning of the head of the noun phrase.[14] Even more subtle linguistic regularities can have a semantic component; as a final example, the preference of some verbs in their arguments can be semantically motivated as in the verb *eat* which takes a noun that has a feature +EDIBLE.

We see two ways humans can learn such semantically motivated patterns depending on whether there is an influence of pre-existing linguistic knowledge. Firstly, as in the case of syntax and phonology, L1 influences the learners' representational space rendering the learnability of such patterns a function of whether the particular regularity is somehow

[14]Throughout the present thesis we refer to the Chinese classifier system without necessarily making any explicit distinction between Mandarin and Cantonese. Doing so is not entirely accurate as the two dialects might differ slightly in this respect –potentially even predicting different patterns. Wherever relevant we do note the dialect that possesses a particular construction otherwise we refer to 'Chinese' to denote the generic classifier construction.

encoded in the L1. In this scenario, English learners of Xhosa are at a disadvantage as the semantic regularities found in the Xhosa noun class system are not reflected in English. Secondly, the semantic patterns which can be learned are somehow universal (akin to the universals in generative syntax) and, thus, shared among speakers of different languages. In this case, learners should not have any problems learning such regularities as they already represent the relevant information. If, for instance, *animacy* is a semantic universal –and it might very well be– then even speakers of languages which do not encode animacy (e.g., English) should not have any problems learning the relevant noun classes in Xhosa.

Classic results on early work in cognitive psychology (Berlin & Kay, 1969; Heider, 1972; Heider & Olivier, 1972) have argued that speakers of different languages share their conceptual space independently of how their respective languages encode different constructions. In a landmark publication, Berlin & Kay (1969) presented the view that cross-culturally the colour terms used in a culture are predictable by the number of colour terms found in that culture (e.g., if a culture has names for three colours then one should be red). Furthermore, in her seminal work, Eleanor Rosch (Heider, 1972; Heider & Olivier, 1972) found that the Dani people in Papua New Guinea while they lacked terms for specific colours, they would, nevertheless, categorise colours in a similar manner as speakers of English. Given these results, she argues (Rosch, 1978) that the human conceptual system is determined by the limits of the world around us and not by language. Such results along with a longitudinal study of creole languages have led Bickerton (1984) to propose that there is a limited number of semantic regularities found in languages around the world all having to do with core conceptual knowledge. Results such as these clearly support the latter of the two hypotheses presented in the above paragraph.

Turning now to semantic implicit learning data consider the system introduced in Fig. 1.5. In behavioural experiments using this system, the participants learn four novel determiners which have a distribution similar to the English articles *the* or *a*. They are, however, told that these determiners also encode the distance of the predicate as shown in the figure. For example, the sentence 'He patted **gi** tiger in the zoo' means that the *tiger* was near to the subject. Normally, in these tasks, the participants are pre-trained on the novel meanings and then they have to indicate their meaning in sentences they read. Unsurprisingly, the accuracy scores in these tasks are at a ceiling or near-ceiling level. In a subsequent testing phase, the participants are asked to generalise to novel nouns (nouns which were not seen during the training phase). The problem lies in the fact that for the participant the determiners can be used interchangeably as they were not told of any cues predicting the upcoming noun. As can been seen in the figure, however, there is also a *hidden* co-occurrence rule controlling the conditional probability of these determiners; one set of near/far determiners are reserved for

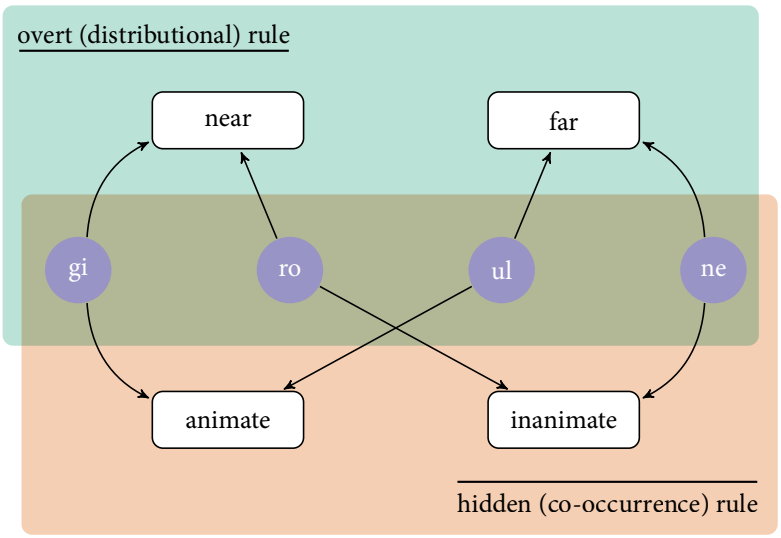


Figure 1.5 The semantically-driven determiner system used in Williams (2005). This has been extended in Paciorek & Williams (2015), Leung & Williams (2014) and Chen et al. (2011) to accomodate different semantic distinctions in the *hidden rule* and different meanings in the *overt rule*. As an analogy to the MNPQ paradigm we also call the *overt* rule distributional and the *hidden* co-occurrence. We recognise that considering the artificial language systems in these experiments, the distributional rule is given we keep this distinction to highlight the commonalities between the two paradigms.

animate nouns while the other for inanimate. The only way for the participants to achieve higher than chance generalisation performance would be to have knowledge of the underlying rule in the bottom part of the figure. There are a few advantages of this paradigm for exploring implicit language learning; firstly, using mainly English lexis alleviates some of the processing bottlenecks of learning an entirely novel system from scratch (cf. DeKeyser, 1995). Secondly, the absence of overt morphophonological markers ensures that participants process something beyond simple surface regularities.

Using this system, Williams (2005) has managed to obtain higher than chance performance in the generalisation task. In two experiments performed on 65 participants (41 and 25, respectively) of various linguistic backgrounds, using the same set of 24 nouns equally divided between animate and inanimate retaining only two for generalisation it was found that participants were able to appreciate the underlying (hidden) regularity without being able to verbalise any rules. The finding of Williams (2005) gave rise to an increased interest in semantic implicit learning revolving around what other semantic distinctions might be learnable (Chen et al., 2011; Leung & Williams, 2012; Paciorek & Williams, 2015) or whether

constructions other than determiner + noun phrases can be learnt (Paciorek & Williams, 2015). Also, interest increased on whether this type of learning extends to speakers of languages other than English (Chen et al., 2011; Leung & Williams, 2014), whether there is L1 influence (Leung & Williams, 2013, 2014) or, finally, what neural structures support this kind of learning (Batterink, Oudiette, Reber & Paller, 2014).

Retaining the original structure, Paciorek & Williams (2015) extended the above paradigm to learning the selectional preferences of verbs instead of determiner + noun collocations. As remarked above, selectional preferences are a type of co-occurrence where some verbs 'prefer' as arguments nouns with a specific semantic feature. For example, as seen in (1.2) and (1.3) the verb *eat* can take as an argument a noun which has the feature [+EDIBLE].

(1.2) He ate pasta.

(1.3) *He ate a laptop.

Apart from this, the experiments reported in Paciorek & Williams (2015) introduce three novel manipulations over the original paradigm in Fig. 1.5. Firstly, instead of focusing on the animate/inanimate distinction, PW examined a semantic distinction between abstract and concrete concepts. The reason for this decision was that it fit better with the experimental paradigm. Secondly, instead of an alternative forced choice task as above, PW employed a false memory task. Participants encounter here novel verb + noun combinations (i.e., nouns not seen previously) asking them whether they have seen this before. Drawing on a vast literature of false memory effects (Brainerd & Reyna, 2005), it was reasoned that if a regularity were abstracted, this would manifest as higher endorsement rates for the grammatical alternatives. We argue later §5.2 that while for the modeller nothing changes between the two test tasks for the behavioural researcher this would alleviate a potential confound. Looking back at the original 2AFC paradigm, participants were presented with two determiners bearing the same meaning; a sceptical participant, therefore, while not having any idea of underlying rules at that stage, might be prompted to look for them during test time. This would result in higher than chance generalisation rates without, however, implicit learning of the rule during the exposure phase. The third issue in which we are interested was the manipulation of the semantic distance between training and test sets. While we will explore this at length in §5.5 consider that during training participants saw a number of chemical elements with the same verb (concrete condition); would they generalise the same to other chemical elements as to other concrete concepts?[15] The data presented by PW was negative as participants showed a

[15]In the actual experiments the participants were trained on 16 abstract and 16 concrete concepts and, subsequently, tested on 32 novel items from each category. More details on the actual stimuli can be found in §C.2.

preference for other chemical elements as opposed to concrete concepts in general. In other words, the semantic distance between the concepts affected the generalisation gradients.[16]

For an implicit learning sceptic, while the above studies demonstrate learning effects without participants being able to verbalise any rules, they use an offline methodology during testing to establish learning. The problem with this is that during testing the participant has the chance to re-evaluate her thinking of the experiment leading to higher than chance results without learning. Further to that, awareness was also measured using post-experimental questionnaires which might lead to false positives (Shanks, 2016). The problem here lies in the fact that a participant might start forming hypotheses during the experiment, but these are either forgotten or not really established in the questionnaire. Leung & Williams (2014) devised another methodology retaining the underlying system of Fig. 1.5. As shown in Fig. 1.6 each trial now instead of offline reading of a short text, consists of a presentation of only the critical NP (e.g., 'gi' dog). The first response the participant has to make is an animacy judgment pressing one of two buttons ('m' for animate and 'c' for inanimate). Subsequently, she has to indicate whether the determiner shown meant near ('m') or far ('k'). Throughout the experiment, the participants were instructed to be as fast as possible hindering any conscious hypothesis formation strategies. Having been trained this way on the system participants started to encounter violation trials which did not follow the rule. If participants had started to appreciate the underlying rule, then a violation trial would result in more processing effort, thus, an increase in the reaction times. Using this methodology, Leung & Williams (2014) managed to find a dissociation between *violation* and *control* trials largely replicating the above results which show semantic influence.

The system presented in Fig. 1.5 can be re-cast as an instantiation of the more common MNPQ paradigm (Braine, 1966). This has the added advantage of exploiting the results of a well-known paradigm permitting us to probe further the processes which underlie semantic implicit learning. Within the MNPQ paradigm, there are four word classes – M, N, P and Q. The appearance of each word is conditioned on two interacting rules; a *distributional* rule which allows *only* M and P words in the first position leaving N and Q words for the latter, and a *co-occurrence* rule which forces P words to follow M and N to precede Q (see Fig. 1.7). Class membership can be arbitrary ranging from complete absence of any predictive cues (Smith, 1969) to specific morphemes signalling the class (Braine, 1987). Relating this to the studies presented above, M and P words are the determiners/verbs and N and Q the nouns. In this case, the *distributional* rule is embedded in a way in the grammatical system (i.e., articles before

[16] At this point, the semantic distance can be thought of encoding the intuition that *oxygen* and *hydrogen* are more similar than *oxygen* and *table*. We present a mathematical definition of semantic similarity in Table 2.4.

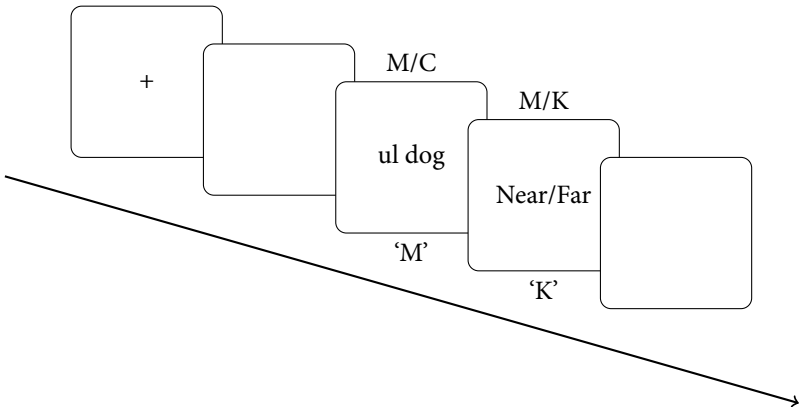


Figure 1.6 A single trial from Leung & Williams (2014). After the presentation of the first NP the participant has to indicate whether the noun is animate or inanimate (pressing 'm' or 'c' respectively). Subsequently, she has to respond on whether the previous determiner meant near or far (pressing either 'm' or 'k'). Correct answers are given below the respective panels.

nouns, verbs before their predicates) so participants do not need to learn it. The *co-occurrence* rule, on the other hand, is the semantic condition between the M/N words and the P/Q.

Classic results from this paradigm conclude that subjects can quickly learn the positions of the word classes (i.e., the *distributional* rule) (Braine, 1963; Smith, 1966), even when another word is embedded between the two-word classes (i.e., as in the related aXb paradigm) (Braine, 1965, 1966). The *co-occurrence* rule, on the other hand, has proved more problematic for learners to uncover. Smith (1969) examined the productions of participants in a free recall task. During the training phase, participants were exposed to the 12 legal word pairs from the MNPQ system. Subsequently, they had to perfectly recall the given list or they would receive the same items again in a different order. Given the design of the system there are four possible types of pairs the participants could produce during recall; pairs which were presented during the exposure phase, novel pairs which follow both rules (GI – grammatical intrusions), novel pairs which follow only the distributional rule (SI – semi-grammatical intrusions), and ungrammatical pairs. Firstly, he found that the proportion of semi-grammatical intrusions was significantly different from chance (i.e., if participants were producing pairs by randomly combining letters) indicating that participants were sensitive to the distributional rule. Secondly, he reasoned that if participants were basing their recall on the co-occurrence rule, then the last type of intrusion to be eliminated would be the GI. However, in a by-subject analysis on the last or next to the last trials, it was found that the obtained GI proportion was smaller

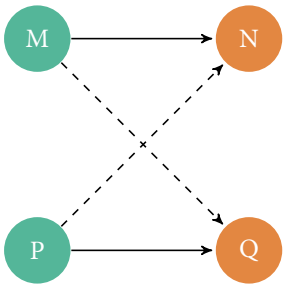


Figure 1.7 The original MNPQ paradigm used in Braine (1966). The rules of the system are as follows; M and P always precede the N and Q words. The *distributional* rule which defines the position of each word category is colour-coded, while the *co-occurrence* rule which defines the conditional occurrence of the words in the second position is shown by the arrows. Solid lines indicate legal transitions respecting both the distributional and the co-occurrence rule while dashed lines show transitions following only the distributional but not the co-occurrence rule. While participants in such experiments can learn the distributional rule, they remain unaware of the co-occurrence rule that N words follow M and P words precede Q, leading to the erroneous endorsements MQ and PN.

than chance, indicating that participants relied either on direct recall of the experimental stimuli or reconstructive processes based on the distributional rule.

Results such as these may seem to contradict those from the semantic implicit learning experiments presented above. There are, however, at least two ways we see the SIL results differing from the early MNPQ. Firstly, within the MNPQ participants learn an entirely novel system of non-words with no overt grammatical relations. The SIL paradigm alleviates this problem by using known words and grammatical constructions. This poses fewer resource demands on the participants. Secondly, in the studies presented above, there were no cues determining class membership. Regarding the second observation, Braine (1987) found that correlating the word classes with natural gender makes the co-occurrence rule more learnable. In one experimental condition participants learned an MNPQ system where half the M and P words were *male* and *female* professions, respectively (the rest being inanimate objects). The P and Q classes in this instance denoted the number of the other classes (i.e., one, two, plural). It was found that participants were not only able to learn the co-occurrence rule (as shown by generation tests) but also that they were able to generalise to unseen nouns. These results bridge the gap between the 'unlearnable' co-occurrence rule in the MNPQ system and the semantic rule in SIL.

Returning to the issue of representations raised in §1.3.1, it remains an open question what kind of information is available under such tasks. Without a doubt, the physical world

provides a number of cues some of which are language-related (e.g., semantics, phonology) and some of which are due to limitations of the experimental procedure (e.g., randomisation of stimuli). Computational modelling aids here in distinguishing what kind of information is abstracted and used in these tasks. If, on the one hand, participants had full access to semantic resources the SIL experiments would be trivialised. On the other hand, if there were no semantic access at all then *probably* the participants would be unable to discover the rules of the system, stressing the uncertainty from *probably*. There are two cues which make us question the relevance of semantics in the present case. Firstly, the low generalisation rates in these experiments, prompt us to think that instead of the semantic distinction, the participants learned a version of the system which while not using semantic information still enables them to achieve higher than chance generalisation rates. The second related reason comes from the acquisition theory put forward by Braine (1987). Braine conjectures there that phonologically determined systems (for example gender systems in Italian or Spanish) should be more easily learnable than German as the salience of the cues is more capitalised in those languages. In §6.2 we explore this idea further by looking at whether the underlying system in SIL experiments is recoverable solely by phonological information.

This Chapter has highlighted some of the main challenges posed by implicit learning in general, and semantic implicit learning, in particular. We have also introduced a set of semantic implicit learning experiments which raise the question as to what kind of semantic knowledge would be responsible for the behavioural results. In what follows, we will describe a particular class of algorithms that gain semantic-like knowledge through distributional analysis of massive linguistic input. Subsequently, in Chapters 3 and 4 we explore the potential of these algorithms to model behavioural data that resemble semantic implicit learning tasks. In Chapter 5 we use the results from the previous Chapters to identify the mechanisms responsible for the behavioural performance in the above tasks, as well as their limitations.

Chapter 2

Distributional Semantics

> Our pockets were full of deng, so there was no real need from the point of view of crasting any more pretty polly to tolchock some old veck in an alley and viddy him swim [...]
>
> Alex DeLarge in *A Clockwork Orange*

2.1 Introduction

In the 1962 dystopian novel *A Clockwork Orange* by Anthony Burgess teenagers speak in a fictional register called *Nadsat*. While its functon is up for debate, from a linguistic point of view it is nothing more than English with some unknown –from the speaker's point of view– words.[1] Interestingly, Anthony Burgess did not provide a key to the unknown words' meanings[2] leaving the interpretation up to the reader. How is the reader, however, able to perfectly restore the intended meaning? On a first level, the part of speech for some of the unknown words can be inferred either morphologically (as in *crasting*) or probabilistically (e.g., *tolchock*). Indeed, parsing this novel through Stanford's Probabilistic Context Free Grammar (PCFG) parser, it correctly identified *crasting* and *tolchock* as verbs.† Still, however, even though we are able to restore the syntax of the sentence and identify it as grammatical, we have no clue what the words used actually mean. A simple strategy we could adopt would be to look at the words around the unknown word (e.g., *deng*) and then infer what other known word we could use in that context. For example, in the phrase 'Our pockets were full

[1] In reality most of the words are borrowed from Russian: *deng* (< Rus. деньги = 'money'), *crasting* (< Rus. красть= 'to steal'), *polly* (rhymes with 'lolly' = 'money'), *tolchock* (< Rus. толчок = 'hit'), *veck* (< Rus. человек = 'person') and *viddy* (< Rus. видеть = 'to see').

[2] A key was provided in the restored version of the book in 1986.

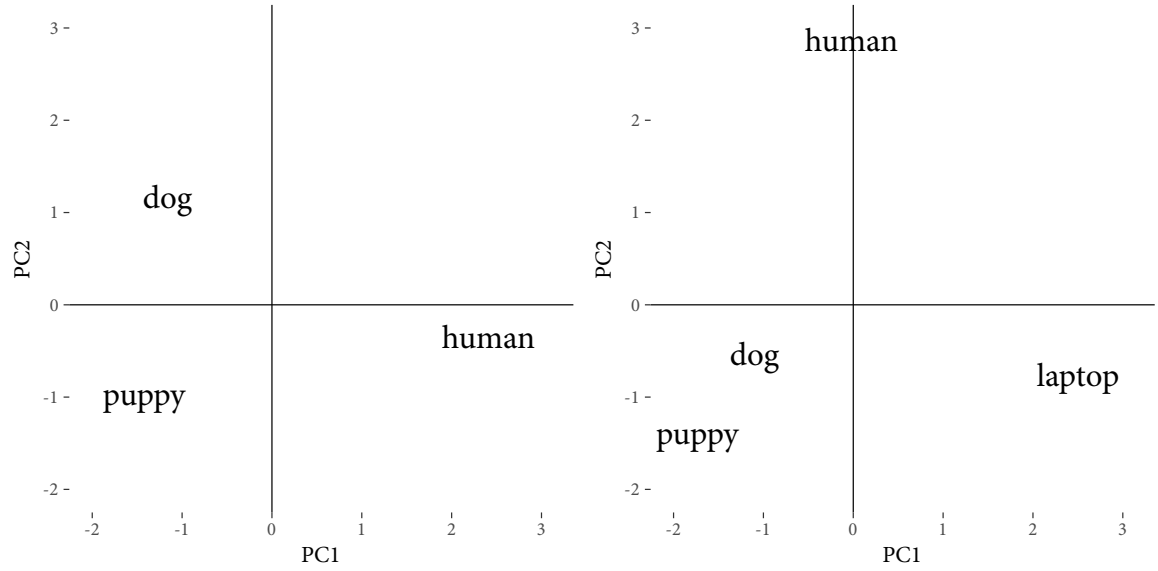
of . . . ', without knowing how the text continues, words such as *money*, *gold* would be probable. In order to carry out this task we need a computational mechanism that is able to gather contextual information for each word, and then use an affinity function to match this novel word to another one from our lexicon.

In this chapter, we explore a certain class of computational models of semantics known as Vector-Space Models, which give words a geometric representation as points in high-dimensional spaces. The derivation of this representation is a function of the distributional characteristics of that word when found in large linguistic corpora. The elements, therefore, of a word vector (or its coordinates in this space) are computed in such a way that they reflect the linguistic environment, in which the corresponding word is found.

Theoretical linguistics have supported the idea that the linguistic environment (the '*context*') provides crucial information about a word's meaning (Firth, 1935, 1957; Harris, 1954). Studies in cognitive psychology (Barclay, Bransford & Franks, 1974; Barsalou, 1987) seem to also support this idea of context-generated meaning, prompting researchers in the field (Andrews, Frank & Vigliocco, 2014; Andrews, Vigliocco & Vinson, 2009; Kintsch & Mangalath, 2011; Landauer & Dumais, 1997) to view the process of *learning 'meaning' from context*, as an abstract computational problem in the spirit of Marr (1982) and Anderson (1990), that the human *semantic memory* attempts to solve by means of some underlying cognitive mechanism. In this formulation, the different Vector-Space Models are but mathematical conveniences for approximating these 'true' cognitive processes.

It is evident, however, that the size of the problem space grows with the number of possibilities of capturing statistical regularities from the context. Deciding what information to include from the context can have an enormous impact on the final word representation and the overall configuration of points in the semantic space. Interestingly, different phenomena, such as *essay grading* (Alikaniotis, Yannakoudakis & Rei, 2016; Landauer, Laham & Foltz, 2003), *summary writing* (Kintsch, Caccamise, Franzke, Johnson & Dooley, 2007), *memory retrieval* (Howard & Kahana, 2002), *similarity judgements*, *word recognition* (Lund, Burgess & Atchley, 1995; Mirman & Magnuson, 2008; Rohde, Gonnerman & Plaut, 2006), *semantic priming* (Jones et al., 2006; Jones & Mewhort, 2007) or even BOLD signals when processing natural language (Mitchell, Shinkareva, Carlson, Chang, Malave, Mason & Just, 2008) are better accounted for by exploiting various aspects of the context.

Following the discussion in §1.3.1, the word representations explored here fall in the word-to-word relation category. However, contrary to word association norms explored in that section, vector-space models provide a more *objective* view of the word associations gathered directly from word usage and not from personal experience.



(a) (b)

Figure 2.1 (a) Example two-dimensional projection of the words *'human', 'dog', 'puppy', 'human'*. We can describe the position of each word in this two-dimensional space with a tuple of $x\,y$ values. We can calculate the distance between two points in that space by using a simple metric such as Euclidean distance. (b) Two-dimensional Principal Components Analysis (PCA) projection of the 300-dimensional neural embeddings (see §2.3.1) vectors for the words *'human', 'dog', 'puppy', 'laptop'*. Relying only on two-dimensional word vectors PCA would not be able to capture that *dog* and *puppy* cluster together as animals while *human* and *dog* cluster together as living things.

2.2 Vector-Space Models

Vector-Space Models represent words as points in a Euclidean space. In the two-dimensional projection in Fig. 2.1,[3] for example, we can represent all the words with a tuple of (x, y) values, which adequately describes their position in that space. The coordinates for the word *dog* in that space are (1.2, .7), of *puppy* (1.4, .8) and so on. The calculation of the distances between the points becomes a trivial matter using simple algebraic procedures. For example, by taking their Euclidean distance (see Table 2.4), which is equivalent to measuring their distance with a ruler, the distance between *dog* and *puppy* in this space is .22 (substantially smaller than the distance between *dog* and *laptop* which is 1.39).

[3]The parameter details for the two-dimensional projections in the Fig. 2.1 are as detailed in §3.3.1

Compressing words in two dimensions can obscure the high-dimensionality of word meanings, thus, rendering 2D projections uninformative. The issue is that there are cases where two words might be considered similar in some aspects, while not in others. To capture richer representations for each word, Vector Space Models construct high-dimensional spaces, where each word w is represented by a vector of scalar values $\mathbf{w} \in \mathbb{R}^n$, where n is the predefined dimensionality of the vector space. While the values of the elements of each vector are meaningless by themselves (Landauer & Dumais, 1997), they store interesting information that makes sense only in relation to other vectors. As an example, we derived 300-dimensional vectors for four words *'dog', 'human', 'puppy', 'laptop'* and subjected them to Principal Component Analysis retaining only the first two principal components shown in Fig. 2.1. In the left panel, we see that the animals *dog* and *puppy* cluster to the left, while *human* to the right. Interestingly, Fig. 2.1b shows that when we include the word *laptop*, the word *human* is pushed together with *dog* and *puppy* along the first principal component, separating, thus, *animate* and *inanimate* concepts. This property can be taken to illustrate the ability of such models to learn more latent characteristics of concepts without being given any supervised information (as, for example, what *animacy* is).

It might be interesting to ask, how the model reached the conclusion that *dog* is closer to *puppy* than it is to *human*. If the statistical distribution of the word among contexts should provide crucial information on its meaning, in which we find the corresponding word, then *dog* and *puppy* should tend to appear in similar contexts. Consider, for example, the sentences,

(2.1) The young dog eats from its plate.

(2.2) If you hate the idea of your dog going into kennels then Bed and Bone may just be the answer for you.

There is a variety of words related to *dog* in these two sentences, such as *eat*, *young*, *kennel* and *bone* to say the least. One way of exploiting context information would be to look at the words that immediately precede and follow *dog*. Applying this logic to (2.1), we capture the fact that the word *dog* can appear in the context of *young* and *eats*. The idea is that another word, such as *cat* that can also be found in a similar context will be considered as related, which will not be the case with an unrelated one (as in §2.2.3 and (2.4))

(2.3) The young $\left[\frac{\text{dog}}{\text{cat}}\right]$ eats from its plate.

(2.4) * The young laptop eats from its plate.

Applying, however, the same strategy in (2.2) this method will fail to capture the relation between *dog* and *kennel* or *bone*. A workaround to this problem would be to widen the

context window, so as to include words, such as *bone*. This captures the intuition that when a word, such as *dog* appears in a phrase/sentence/text, we expect a related word, such as *bone* to be found in the same text, regardless of the position of *dog* (Landauer & Dumais, 1997). Computationally, exploiting this kind of information from the context would require gathering statistics not between words, as before, but between words and texts. If therefore, two words seem to co-occur across documents then there must be some relation between them.

Theoretical linguistics (de Saussure, 1916) and more specifically, *lexical semantics* (Cruse, 1986) have used the terms *paradigmatic* and *syntagmatic* to describe the above relationships. These two ways to encapsulate context can have a significant impact on the ability of each model to approximate semantic representations. The reason for that is that a *syntagmatic* model will tend to favour word pairs that have more of an associative relation (i.e. *dog* and *kennel*), while a *paradigmatic* will place greater weight on a more *semantic* relationship such as the one between *dog* and *cat* (Sahlgren, 2008).

2.2.1 Syntagmatic Models

Assuming that each of the examples in 2.5–2.8 is a document or the *context*, in which words such as *modem* can be found (Ruge, 1992), we can construct a *term-document* matrix $\mathbf{M} \in \mathbb{R}^{|\mathcal{V}|\times|C|}$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ is the set of all the different word types and $C = \{c_1, c_2, \ldots, c_m\}$ is the set of different texts (by $|\cdot|$ we denote the cardinality of the set). Each cell M_{ij} in this matrix is a count of how many times the word i occurred in the context j. If we substitute v_1 = 'linux' and c_1 = (2.5) then $M_{11} = 3$. Assuming a constant vocabulary, the terms in this matrix grow *linearly* with the number of documents.

(2.5) modem the steering linux. modem, linux the modem. steering the modem. linux!

(2.6) linux; the linux. the linux modem linux. the modem, clutch the modem. petrol.

(2.7) petrol! clutch the steering, steering, linux. the steering clutch petrol. clutch the petrol; the clutch.

(2.8) the the the. clutch clutch clutch! steering petrol; steering petrol petrol; steering petrol!!!!

Table 2.1 shows the resulting matrix derived from the sentences (2.5) and (2.8), from which the 4D-vectors shown in (2.9) are extracted. The similarity / distance between the vectors can be found using any metric from Table 2.4, which returns a scalar $x \in [-1, 1]$, where a value of -1 means that the two vectors have opposite directions (not similar at all), and a value of 1

Table 2.1 Term × Document Matrix from the sentences (2.5) and (2.8) prior to any normalisation procedures.

Word	Document 1	Document 2	Document 3	Document 4
linux	3	4	1	0
modem	4	3	0	1
the	3	4	4	3
clutch	0	1	4	3
steering	2	0	3	3
petrol	0	1	3	4

[1] Any sentence, phrase or longer chunk of text can be considered a document in these models. It is worth noting that in this framework the order of the word in the sentence does not play a significant role.

are collinear (identical). A drawback of this geometric approach is that the similarity between two word-vectors $sim(\mathbf{a}, \mathbf{b})$ is *commutative* (i.e. $sim(\mathbf{a}, \mathbf{b}) = sim(\mathbf{b}, \mathbf{a})$), violating, thus, the metric axioms put forward by Tversky (1977) (also, Griffiths et al., 2007), in that similarity between two concepts can be *asymmetric*. Using, however, a distance measure, such as the inverse squared euclidean (Table 2.4), the similarity between *linux* and *modem* in this 4D space is .2, whereas between *linux* and *petrol* is .02.

$$\mathbf{v}_{\text{linux}} = \begin{bmatrix} 3 \\ 4 \\ 1 \\ 0 \end{bmatrix}, \mathbf{v}_{\text{modem}} = \begin{bmatrix} 4 \\ 3 \\ 0 \\ 1 \end{bmatrix}, \cdots, \mathbf{v}_{\text{petrol}} = \begin{bmatrix} 0 \\ 1 \\ 3 \\ 4 \end{bmatrix} \tag{2.9}$$

A closer look at Table 2.1 reveals that the vector of *the* receives high counts across all documents, implying that is relevant to all contexts. This happens because a small number of words, such as *the* are much more frequent than most of the words in a corpus (Baayen, 2001; Zipf, 1935, 1949), and are, thus, *expected* to have high frequencies across contexts. From an information theoretic perspective, however, *expected* events have lower information content than unexpected ones (Shannon, 1948). It is desirable, therefore, to reweigh this matrix using some *normalisation* function (Table 2.3), so as to better approximate the relationships between terms and documents. For example, using the popular *tf-idf* normalisation procedure (Sparck Jones, 1972) each term in the vocabulary receives a high weight if it appears frequently

Table 2.2 Term × Document Matrix from the sentences (2.5) and (2.8) after having applied *tf-idf* normalisation.

Word	Document 1	Document 2	Document 3	Document 4
linux	0.0664	0.0959	0.0192	0
modem	0.0885	0.0719	0	0.0205
the	0	0	0	0
clutch	0	0.0240	0.0767	0.0616
steering	0.0443	0	0.0575	0.0616
petrol	0.0221	0	0.0575	0.0822

Table 2.3 Common distributional semantic vector normalisation procedures

Length	$M_{ij} = \frac{M_{ij}}{\sqrt{\sum_{k=1}^{D} M_{ik}^2}}$
Orthographic Frequency	$M_{ij} = \frac{M_{ij}}{\text{freq}(M_i)}$
Correlation	$M_{ij} = \frac{TM_{ij} - \sum_k M_{ik} \cdot \sum_l M_{lj}}{\sqrt{\sum_k M_{ik} \cdot (T - \sum_k M_{ik}) \cdot \sum_l M_{lj} \cdot (T - \sum_l M_{lj})}}$ $T = \sum_k \sum_l M_{lk}$
Entropy	$M_{ij} = \frac{\log(M_{ij} + 1)}{\mathbb{H}(i)}$ $\mathbb{H}(i) = -\sum_k \frac{M_{ij}}{\sum_k M_{ik}} \log\left(\frac{M_{ij}}{\sum_k M_{ik}}\right)$
Tf-Idf	$M_{ij} = M_{ij} \times \log \frac{K}{\sum_k^K M_{ik}}$
Sub-sampling[1]	$P(w_i) = 1 - \sqrt{\frac{t}{\text{freq}(w_i)}}$

[1] Sub-sampling defines the probability of a word being dropped from training depending on its frequency.

Notes: M is the TERM × CONTEXT matrix; D is the dimensionality of the context; freq(·) the frequency of a word in the corpus and t a threshold parameter (see text in §2.3.1).

in a specific document but is rare otherwise (Table 2.2), capturing the intuition that a frequent word does not carry interesting semantic meaning (Rohde et al., 2006).

Latent Semantic Analysis

From an informal experiment on the ukWaC corpus (Baroni, Bernardini, Ferraresi & Zanchetta, 2009) out of the 55949 times the word *boat* occurs in the corpus, only 843 times it occurs

Table 2.4 Common vector similarity measures

Euclidean Distance[1]	$d(\mathbf{w}_1, \mathbf{w}_2) = \sqrt{\sum_{i=1}^{n}(w_{1_i} - w_{2_i})^2}$
Inverse Squared Euclidean	$sim(\mathbf{w}_1, \mathbf{w}_2) = \frac{1}{\sum_{i=1}^{n}(w_{1_i} - w_{2_i})^2 + 1}$
Cosine	$sim(\mathbf{w}_1, \mathbf{w}_2) = \frac{\mathbf{w}_1 \cdot \mathbf{w}_2}{\|\mathbf{w}_1\| \|\mathbf{w}_2\|}$

[1] The Euclidean metric measures the *distance* between two vectors, returning a value that decreases with similarity. Following Rohde et al. (2006), we invert this function and add 1 to the denominator so that the value is always bounded between 0 and 1.

with the word ship. This problem is common with near-synonymous words as they tend not to co-occur in the same document (Clark, 2015) and as a consequence, by calculating the distance between their vectors, the words would appear as markedly dissimilar. Indeed, the cosine similarity between *ship* and *boat* based on raw counts is only .01. How can Vector-Space models capture the intuition that the words *boat* and *ship* are almost synonymous? The answer is trivial, in that these two words might not co-occur but rather they consistently occur in the context of other words, such as *harbour*, *pier*, *sea* etc. (Landauer & Dumais, 1997; Lemaire & Denhière, 2006). The idea that two words can be associated through a third one is well-grounded on psycholinguistic studies of *mediated* priming (Balota & Lorch, 1986; de Groot, 1983; McKoon & Ratcliff, 1992; McNamara, 1992; McNamara & Altarriba, 1988), in which words such as *lion* can be associated with *stripes* through the word *tiger*.

Viewed this way, learning latent semantic relationships requires learning of higher-order (indirect) co-occurrences (Lemaire & Denhière, 2006) while the original TERM × DOCUMENTS matrix offers first-order co-occurrences (direct). Discovering these relationships, therefore, requires a good deal of inductive inference. Landauer & Dumais (1997) argue that a mathematical approximation to this problem of induction would be to lower the dimensionality of the original matrix. Vozalis & Margaritis (2003) and Turney & Pantel (2010) offer a way of capturing how this works, viewing it as a *data sparsity* problem. The original vector of *boat* has lots of zeros in the contexts where it *could* have positive values (i.e. in the contexts where *ship* appears). Compressing the dimensionality compensates for this lack of data, by enforcing greater correspondence between terms and contexts. Similarly, Shepard (1987) has argued that lowering the dimensionality of sparse representations leads to uncovering latent information.

Deerwester, Dumais, Furnas, Landauer & Harshman (1990) have explored the use of Singular Value Decomposition (SVD) (see, also, Manning & Schutze, 1999) in performing dimensionality reduction. The main idea is that we can factorize the original matrix $\mathbf{M}$ as the product of three new matrices $\mathbf{U}\Sigma\mathbf{V}^{\mathsf{T}}$. Roughly speaking, the matrix $\mathbf{U}$ holds the word vectors,

while the matrix $\mathbf{V}$ holds the document vectors. The matrix Σ, on the other hand, holds the weights that rate the importance of the vectors. It can be proved that if the weights (or more formally *singular values*) are permuted in such a way as to be ranked in decreasing order, then truncating all but the first k, returns the best k rank approximation of the original matrix.

$$\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^{\top} \tag{2.10}$$

$$\hat{\mathbf{M}}_k = \hat{\mathbf{U}}_k\hat{\Sigma}_k\hat{\mathbf{V}}_k^{\top} \tag{2.11}$$

Using the final $\hat{\mathbf{U}}_k$ matrix, we can use any similarity measure from Table 2.4 to find the new similarity between *ship* and *boat*. In an Latent Semantic Analysis (LSA) model trained on the British National Corpus, using $k = 300$ the new similarity rises to cos('ship', 'boat') = .509.

SVD is, however, computationally a very costly operation. Moreover, since its complexity depends on the number of dimensions we retain *after* truncating the original matrices[4] it is quite difficult to perform experiments where $k > 400$ (on a modern machine). A similar idea that avoids the SVD performance bottleneck is Random Indexing (Kanerva, 2009; Kanerva, Kristoferson & Holst, 2000; Sahlgren, 2006; Sahlgren, Holst & Kanerva, 2006). This requires viewing the whole processes of syntagmatic models somewhat differently; Consider each LSA *context* as a unit vector as in Eqs. (2.12) to (2.14) so that each context would be orthogonal to the others. Each word is assigned a zero vector $\mathbf{0}$ of dimensionality equal to the number of contexts. Each time a word is encountered in a context, this context vector is added to the word vector (i.e., incrementing by one the corresponding cell in the corresponding row). Carried over the whole corpus this operation would ultimately yield the same TERMS × DOCUMENTS matrix as in Table 2.1.

$$\hat{c}_1 = \left(1, \underbrace{0, 0, \ldots, 0}_{D-1}\right)^{\top} \tag{2.12}$$

$$\hat{c}_2 = \left(0, 1, \underbrace{0, \ldots, 0}_{D-2}\right)^{\top} \tag{2.13}$$

$$\hat{c}_m = \left(\underbrace{0, 0, \ldots, 0}_{D-1}, 1\right)^{\top} \tag{2.14}$$

[4]The time complexity of SVD is $O(\min\{mn^2, m^2n\})$, where m is the number of words in $\mathcal{V}$ and n the number of documents. Its complexity, therefore, will grow *exponentially* with the number of dimensions.

The idea of Random Indexing stems from an observation by Hecht-Nielsen (1994) (which was later further developed in Sahlgren, 2006) that one can approximate orthogonal vectors such as those in Eqs. (2.12) to (2.14) by taking *nearly*-orthogonal random vectors of much lower dimensionality. The difference of these *nearly*-orthogonal vectors is that instead of initializing all the elements but one to zero, we pick element positions randomly[5] and set them to either 1 or −1. While this might introduce spurious correlations given a small corpus or a small vector size, it can be shown (Kanerva et al., 2000) that as the sizes of the corpus and the vector increase, similar words will result with similar vectors. Using, therefore, *random* vectors of dimensionality d such that $d \ll c$ and simply adding them to the word vector every time the word is encountered in the corresponding context would ultimately yield a matrix similar to the one after SVD.

2.2.2 Paradigmatic Models

In the Hyperspace Analogue to Language (HAL) the representation of a word is again a function of the contexts it occurs in. There is, however, a fundamental difference between HAL and LSA in the way they encapsulate the notion of context. In LSA two words were considered as occurring in the same context if they occurred within the same text, paragraph or sentence, regardless of their relative position. HAL, on the other hand, exploits the immediate neighbourhood around the target word. Two words, therefore, are more similar if they occur in the same position in a phrase rather than if they occur in the same sentence (see (2.1) and §2.2.3).

To carry this operation HAL constructs a TERM × TERM matrix $\mathbf{M} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ is again the vocabulary and $|\cdot|$ the cardinality of the set. The value of each cell M_{ij} is simply a count of the times word j occurs in the context of word i. To define this context, HAL looks at the words preceding and following the word i within a specified window w. If word j occurs within the distance w from the word i then the value M_{ij} is incremented as a function of that distance. This last step aims to capture the fact that words closer to the target word would arguably be more informative. For example, taking $w = 5$ and $d(i, j) = 1$ (i.e. j is adjacent to i and the context window is 5), then using a function that places larger weights to words closer to the target such as the *linear ramp* function $M_{ij} = (w - d + 1)$ where w is the width of the window and d the distance of word i from word j, $M_{i,j} = 5, M_{i,j+1} = 4, \ldots, M_{i,j+w-1} = 1$.

Specifically, consider the sentence:

(2.15) The old dog chases the angry cat. The mouse and the parrot observe.

[5]Each choice can be considered a Bernoulli trial with the parameter p controlling the bias towards *denser* versus *sparser* vectors. This parameter p can be left as a hyperparameter to the model.

Table 2.5 Example of HAL co-occurrence matrix of the sentence *The old dog chases the angry cat. The mouse and the parrot observe.* The parameters used were: Window (B5A5) and linear ramp weighting scheme.

	the	*old*	*dog*	*chases*	*angry*	*cat*	*mouse*	*and*	*parrot*	*observe*
the	8	5	4	3	6	4	7	5	7	5
old	3	0	5	4	2	1	0	0	0	0
dog	5	0	0	5	3	2	0	0	0	0
chases	7	0	0	0	4	3	1	0	0	0
angry	5	0	0	0	0	5	3	2	0	0
cat	7	0	0	0	0	0	4	3	1	0
mouse	4	0	0	0	0	0	0	5	3	2
and	5	0	0	0	0	0	0	0	4	3
parrot	0	0	0	0	0	0	0	0	0	5
observe	0	0	0	0	0	0	0	0	0	0

Note that the values in this table are prior to any row normalisation method

This passage contains 10 different words, which we can put into a 10×10 matrix. Table 2.5 shows the resulting co-occurrence matrix using the *linear ramp* weighting function introduced above and a window of B5A5 (5 before the target word and 5 after). A characteristic of this table is that it is not symmetric (i.e. $M_{ij} \neq M_{ji}$). This happens because there is different information in the forward ('the old') and the backwards ('old the') association. In order to take into account both kinds of information, HAL concatenates the row vector and the column vector defining, thus, a vector of dimensionality $2|\mathcal{V}|$ (2.16).

$$\mathbf{w}_1 \in \mathbf{M}^\top \| \mathbf{w}_1 \in \mathbf{M} = \left.\begin{bmatrix} 8 \\ 5 \\ \vdots \\ 0 \\ 0 \end{bmatrix}\right\} \in \mathbb{R}^{20} \tag{2.16}$$

As before, the use of raw co-occurrence statistics weighs more *uninformative* words, such as *the*, giving them denser vectors. In HAL simulations, this frequency bias is normally dealt with by using either the *length* (Burgess & Lund, 2000) or the *orthographic frequency* normalisation procedure (Buchanan, Westbury & Burgess, 2001; Shaoul & Westbury, 2006).

A final problem in HAL is the high dimensionality of the resulting vectors. Performing any kind of operation such as finding the cosine distance between two vectors of dimensionality equal to the size of the vocabulary (i.e. in normal cases between 100000-150000 words) can be very cumbersome. While several ways have been proposed to deal with this issue, amongst them to use SVD (Rohde et al., 2006), in practice, retaining a number of columns with the greatest variance (eliminating words that occur very often or very rarely) seems to provide the best results.

An interesting derivation of the original HAL model is the Correlated Occurrence Analogue to Lexical Semantics (COALS) (Rohde et al., 2006), which frames the normalisation procedure differently. In the original HAL model the raw word co-occurrences are of primary interest. A word i is going to be considered close to a word j if they both occur systematically in the context of other words. COALS, however, asks whether word i occurs with some word j more systematically than it does with other words. Following the correlation normalisation procedure described in Table 2.3 each cell is going to have a value between -1 and 1. A positive correlation in this context means that a word j is more likely to occur in the context of some word i. Following this logic, COALS places words closer if these words are more likely to occur in similar contexts rather than in dissimilar.

In addition, the above discussion about the Random Indexing procedure is also relevant in the case of paradigmatic models. The difference in the present case is that each *word* is assigned a random vector which is *nearly*-orthogonal to all the others. Every time, therefore, a context word appears in the window of the target, its vector is added to that of the target, yielding a $|\mathcal{V}| \times D$ matrix, where D is the predefined dimensionality of the random vectors.

2.2.3 Bound Encoding of the Aggregate Language Environment (BEAGLE)

The discussion so far has assumed a clear distinction between paradigmatic and syntagmatic models. However, *hammer* is close to both *nail* (syntagmatic relation) and *screwdriver* (paradigmatic relation), something that neither model discussed above, is unable to capture. The BEAGLE (Jones et al., 2006; Jones & Mewhort, 2007) was specifically designed to overcome this limitation by learning simultaneously a HAL-like and an LSA-like representation. The final *composite* representation is formed by learning a *context* (LSA-like) and an *order* (HAL-like) vector independently and then linearly combine them.

More specifically, the first time a word is encountered in the corpus, it is assigned a random 'environmental' vector sampled from a normal distribution, representing the word's structural

characteristics.[6]

$$\mathbf{e}_w \sim \mathcal{N}(0, \frac{1}{\sqrt{D}}), \mathbf{e}_w \in \mathbb{R}^D \tag{2.17}$$

Although the *contextual* representation of each word in BEAGLE is basically an LSA vector, for performance reasons, BEAGLE employs a Random Indexing procedure summing over all the environmental vectors found in the sentence. Since vector addition is *commutative*, then words appearing in the similar contexts, regardless of word order, will have similar *contextual* vectors. For example, in the word 'plate' will be reflected in the vectors of both *dog* and *cat* as -geometrically- adding the vector of the word 'plate' will perturb the vectors towards that direction. As already noted, the advantage of this approach is that there is need for a separate dimensionality reduction phase as vector addition does not affect the dimensionality of the vector.

$$\mathbf{c}_i = \sum_{j=1}^{N} \mathbf{e}_j, i \neq j \tag{2.18}$$

where N is the length of the sentence.

In order to capture order information about each word BEAGLE avoids TERM × TERM matrices as dimensionality reduction techniques can be both costly and under-informative (Rohde et al., 2006). Influenced by studies on associative memory (Murdock, 1982, 1992, 1993) BEAGLE uses non-commutative *circular convolution* (Plate, 1995, 2003) to combine word vectors, taking into account their order. *Circular convolution* is a method for forming a new vector that is not related to either argument vector. Using a non-commutative version of this method (Plate, 1995), BEAGLE surpasses the problem posed by the commutativity of vector addition, forming a word representation which retains the order information of the context.

In detail, BEAGLE collects all the possible n-grams for a given window around each word (usually three before and three after) as in (2.20)[7], circularly convolves the environmental vectors of those words and finally summing over all of them. Since circular convolution (as opposed to regular convolution) does not alter the dimensionality of the argument vectors, the resulting 'order' vector can be directly combined with the 'contextual' representation, reflecting also the order of the n-grams in the sentence.

(2.19) The quick brown fox

[6] In other words, representing that each word is a unique string.

[7] The Φ in (2.20) is another environmental vector for the target word used only in the derivation of the 'order' representation.

$$\mathbf{o}_{\text{brown}} = \begin{bmatrix} \Phi & \circledast & e_{\text{fox}} & + & & & \\ e_{\text{quick}} & \circledast & \Phi & + & & & \\ e_{\text{quick}} & \circledast & \Phi & \circledast & e_{\text{fox}} & + & \\ e_{\text{the}} & \circledast & e_{\text{quick}} & \circledast & \Phi & + & \\ e_{\text{the}} & \circledast & e_{\text{quick}} & \circledast & \Phi & \circledast & e_{\text{fox}} \end{bmatrix} \tag{2.20}$$

Having obtained *order* and *context* vectors for each word in the corpus, BEAGLE sums over the two to obtain a composite 'memory' representation (2.21) which will encode both sources of information,

$$\mathbf{m}_i = \mathbf{c}_i + \mathbf{o}_i \tag{2.21}$$

2.3 Predictive models

2.3.1 Neural Embeddings

The syntagmatic and paradigmatic models presented above are based on 'counting'. They form semantic representations by counting the number of times the target word occurs in a certain context. Despite their differences in conceptualising context (either as a window around the target or the document in which it appears) or in 'counting' (either by incrementing a cell in the co-occurrence matrix or by adding vectors), these models adhere to the same logic (Baroni, Dinu & Kruszewski, 2014). Higher-order co-occurrence patterns can be recovered using 'inference' mechanisms such as dimensionality reduction.

Take the vector for the word *cat* in Table 2.5. Each element in that row defines the frequency with which the word *cat* appears in a window containing any of the other words.[8] Dividing each cell in the matrix by the sum of each row gives us the *probability* the target word will appear in a particular context. For example, $M_{7,7} = 4$, while $\sum_{\forall x \in M_{7\star}} x = 15$; the probability, therefore, that the word *cat* will appear in a context which contains the word mouse is $4/15 \approx .36$. The same logic can extend in Table 2.1 where the context is defined not as a word but as an entire document. Viewed this way, a word vector is nothing more than the conditional probability distribution of a word given the contexts in which it appears. We use this description to stress some problems associated with the representations produced by the above systems and ways to fix them.

Discrete probability distributions, like the one obtained above, suffer from some problems related to the *curse of dimensionality*. Simply by looking at Table 2.5, we see that 60% of the

[8] As remarked above, these counts are reweighed by the distance between the cue and the target words. However, the same argument holds.

values are zeros; we have already argued that dimensionality reduction can be used to recover latent regularities from such sparse matrices in an approximate way. However, changes in the raw probability distributions can have a drastic impact on techniques such as the SVD used to lower the dimensionality. This observation is important when deriving semantic representations from different corpora; ideally, we would like the representations to be similar across corpora.

A solution to this problem would be instead of gathering statistics to form a *discrete probability distribution* to find a *continuous probability function* of lower dimensionality to describe the high-dimensional discrete distribution. Concretely, let $\mathbf{y}_i$ be the discrete probability distribution for word i. Also, let $\mathbf{x}_i$ be a lower dimensional description of $\mathbf{y}_i$ such that the transformation/translation $\mathbf{A} \cdot \mathbf{x}_i + \mathbf{b}$ yields $\mathbf{y}_i$. The advantage of this approach is that $\mathbf{x}$ is a *smooth function* of the feature values in $\mathbf{y}$. Therefore, small changes in $\mathbf{x}$ will induce small changes in the probability. We now examine a class of models which form semantic representations by predicting the contexts (Bengio, Ducharme, Vincent & Jauvin, 2003; Collobert & Weston, 2008; Mikolov, Corrado, Chen & Dean, 2013b; Mikolov, Sutskever, Chen & Corrado, 2013c; Turian, Ratinov & Bengio, 2010) the words appear in, learning both the *smooth* lower-dimensional representation $\mathbf{x}$ as well as the transformation matrix $\mathbf{A}$.

Consider that we assign to each word in a vocabulary $\mathcal{V}$ a one-hot vector (a vector where all the elements are zero except for one) such that every word is orthogonal to each other. In this case, we construct a $|\mathcal{V}| \times |\mathcal{V}|$ diagonal matrix where every word *type* can be substituted by a word vector. Let also, $\mathbf{L} \in \mathbb{R}^{D \times |\mathcal{V}|}$ be an –initially random– matrix such that $\mathbf{x}_i = \mathbf{L} \cdot \mathbf{w}_i$, $\mathbf{x}_i \in \mathbb{R}^D$. The goal of this system is to learn the parameters $\mathbf{L}$, $\mathbf{A}$ and $\mathbf{b}$ such that $\sigma(\mathbf{A} \cdot (\mathbf{L} \cdot \mathbf{w}_i) + \mathbf{b}) = \mathbf{y}_i$. In other words, we want to learn a *transformation* from the sparse one-hot representation to some latent structure and another *transformation* from that latent representation to the discrete probability distribution.

Fig. 2.2 illustrates what this class of models is trying to achieve. Each word in the input is assigned an one-hot vector (here, for example, the vector for cat is $[0, 0, 0, 1, 0, 0]$); multiplying this vector with the *word embedding matrix* $\mathbf{L}$ yields a latent (hidden) representation. Initially, $\mathbf{L}$ is a uniform random matrix ($\mathbf{L} \sim U[-0.5, 0.5] \in \mathbb{R}^{D \times |\mathcal{V}|}$). This latent representations –the semantic vector– is then multiplied by the *context embedding matrix* $\mathbf{A}$ yielding the cosine similarity of the target word to every other word in the vocabulary (see Table 2.4). Using a non-linear function such as the *softmax* we obtain the probability distribution $\mathbf{y}_i$ of the word i.

$$\mathcal{S}(\mathbf{x}) = P(\mathbf{y}|\mathbf{x}) = \frac{e^{\mathbf{x}^\top \mathbf{w}_j}}{\sum_{k=1}^{K} e^{\mathbf{x}^\top \mathbf{w}_k}}, \quad \forall j \in \mathbf{y} \tag{2.22}$$

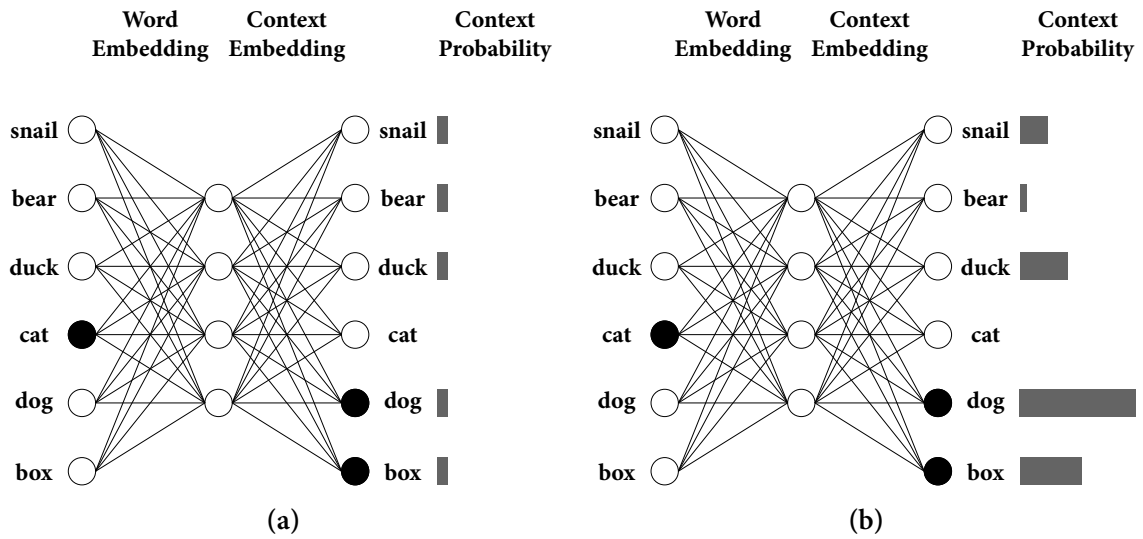


Figure 2.2 Semantic learning as context maximization. The gray bars on the right show the probability that a certain word will appear as context to the target. (a) The probability the network assigns to each possible context (initially uniform across all contexts). (b) The probability at the end of the training phase as the network has discovered which are the possible contexts for each word.

Learning in these systems takes place in a similar manner as in the connectionist networks described in §1.3. The network takes as input an n-gram where the middle word is the target, and the rest provide the context. The one-hot representation of the target word is presented to the input layer of the network, multiplied by the matrix **L**. This intermediate representation is, again, multiplied by the context matrix **A**, 'translated' by the bias vector **b** followed by a point=wise application of non-linearity. The words in the context provide the teaching pattern as the goal of the network is to maximise the probability of these elements in the output layer (see (2.23). The network, then, compares its predictions to the teaching patterns (essentially, whether its predictions were close to 1 on the corresponding elements in the output layer) computing the error gradients, backpropagating the errors and updating the weights in **L**, **A** and **b**.

$$\arg\max_{\theta} \sum_{\mathbf{x} \epsilon Text} \left[\sum_{i \epsilon \mathbf{y}} \log p(\mathbf{y}_i | \mathbf{x}; \theta) \right] \tag{2.23}$$

Computing, however, the Jacobian matrix for every element of the hidden layer to every element in the output layer $\mathbf{J} = \nabla E(w_i) = \left[\frac{\partial \mathbf{E}}{x_1} \dots \frac{\partial \mathbf{E}}{x_D} \right]$, where **E** is our cost function, is a *very* costly operation as it depends on the size of $\mathcal{V}$. An approximation to this operation

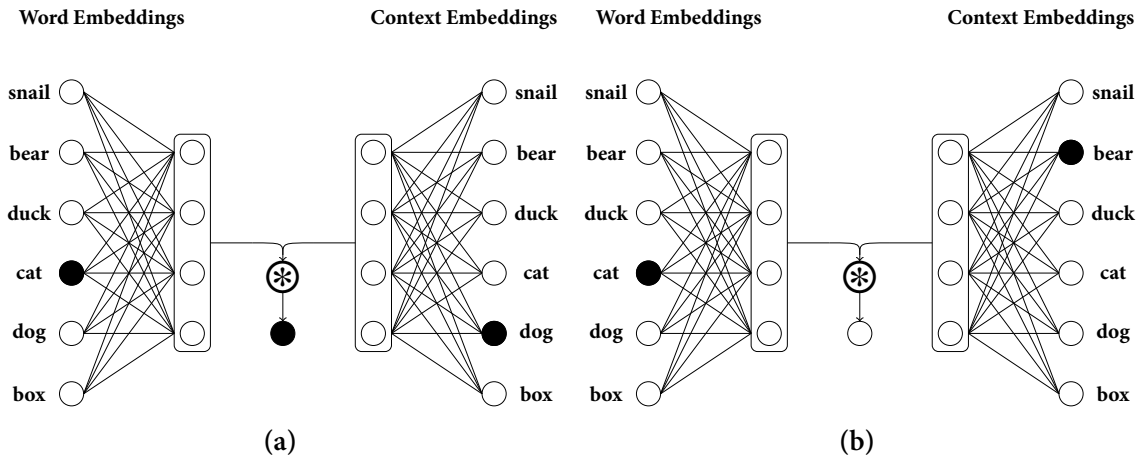


Figure 2.3 Neural Embeddings Learning by Negative Sampling. Two independent neural networks, one for words and one for contexts, are initialised. The architecture of the two networks is exactly the same in terms of layer sizes. The target words enter the left network and the context words the right. The two networks independently compute the hidden representations and then the two hidden layers are fused together into one. The final hidden layer is used to predict whether the example was a positive one (in which case the output unit is 'on') or a negative one (output is 'off').

can be achieved via negative sampling (Gutmann & Hyvärinen, 2012).[9] Fig. 2.3 shows this approximation; at first, both the word and its *true* context enter the network from separate input layers. Subsequently, they are combined using element-wise multiplication. In the sigmoid output layer the network has to predict whether the context word was a *true* context (Fig. 2.3a) or a random word from the vocabulary $\mathcal{V}$ (Fig. 2.3b). Increasing the number of random *false* contexts can significantly improve context prediction in small corpora (Collobert, Weston, Bottou, Karlen, Kavukcuoglu & Kuksa, 2011; Mikolov et al., 2013c).

There are a few advantages of the neural embeddings approach of *traditional* distributional semantics models. Firstly, the number of parameters to be learned grows linearly with the size of the vocabulary rather than exponentially. More specifically, in any HAL-like model (where any word can potentially be context to another) the number of parameters is $|\mathcal{V}|^2$ where $|\mathcal{V}|$ is the size of the vocabulary. In simpler terms, in a corpus of $1.2e^5$ distinct word types (roughly the size of the British National Corpus taking into account words which appear five times or more), the number of parameters to be learned is $1.44e^{10}$ (ca. 14 billion parameters). On the

[9]Predicting the word by its context has come to be known as the Continuous Bag-of-Words model. Alternatively, one can predict the context from each word (which called the Skip-Gram model). In practical terms, the two models yield *very* similar results, and CBoW models are more suitable for larger corpora (Mikolov, Chen, Corrado & Dean, 2013a). For the rest of this thesis, we are going to use the CBoW model solely.

other hand, in the 'predict' approach the number of parameters is substantially reduced to the order of a few million. The learnable parameters are the two matrices **L**, **A** and the bias vector **b**. Since the dimensionality of each matrix is equal to $|\mathcal{V}| \times D$, the number of learnable parameters for a model where $D = 300$ is $2(|\mathcal{V}| \times D) + |\mathcal{V}| = 7.212e^7$. Secondly, the issue of 'online' learning becomes trivial in this context. Continuing the learning procedure beyond the training corpus, or learning new words requires complex algorithms not guaranteed to yield perfect results in systems where an SVD-like algorithm is used to reduce the dimensionality of the original word vectors (Brand, 2006). The above system, however, does not suffer from this problem as it does not have a separate dimensionality reduction step.

If the connectionist framework presented in §1.3 provides a –roughly– accurate account of human learning, then the above system, which is, in essence, a connectionist network, might have implications for semantic learning. Distributional semantics models, while quite successful in capturing relations between words in a large text, as far as psychological and computational tasks are concerned, have been criticised as inadequate models of semantic memory. The reason behind this is that either use mechanisms such as SVD which seem implausible to be implemented by biological structures or their space requirements are simply *too* large. Despite the problems associated with connectionist networks discussed above, neural embeddings sidestep both these issues prompting us to ask what could such models say about semantics.

Let us take another look at what these networks try to accomplish; as the system encounters each word in the corpus, it is trying to predict its permissible contexts. A mismatch in this operation would result in difficulty in processing and higher error gradients when attempting to 'fix' the mistakes. Although evident from Fig. 2.2, it is important to stress that while words enter the network in the linear order in which they appear in the sentence, the context prediction does not necessarily respect this linearity. That is, if the network parses (2.15), the word *cat* is going to enter the network after the word *dog*. However, when predicting the permissible contexts for the word *dog*, the network will not know whether the word *cat* precedes or follows the word *dog* in the sentence. The words in a phrase enter the network as clusters of mutually predicting stimuli. This account that humans instead of predicting upcoming events predict coalesced 'communities' of mutually predicting stimuli from sequential input has been favoured recently by Schapiro, Rogers, Cordova, Turk-Browne & Botvinick (2013) and Schapiro, Turk-Browne, Norman & Botvinick (2015). These authors have shown through behavioural testing, functional magnetic resonance imaging (fMRI) and computational simulations (Schapiro, Turk-Browne, Botvinick & Norman, 2016) that such statistical contingencies capture better the underlying learning mechanisms involved in statistical learning than transitional probabilities.

What does the network learn during this process? The quantity of interest is, without a doubt, the *word embedding matrix* **L** which contains the semantic representations for all the words in the vocabulary (since each column vector is, essentially, the low-dimensional description of the probability distribution). During this *prediction* procedure, the network sets its weights in such a way such that similar words share similar configuration of weights. From a machine learning point of view, this should be no surprise; similar probability distributions *should* share similar lower-dimensionality representations. From a cognitive point of view, however, this is far from trivial. Within this description, learning semantics is simply a by-product of the attempt to minimise the prediction error of incoming information.

The above formulation is, unfortunately, marred by a few issues; firstly, the above formulation defines semantics *solely* via the associations words form with each other. We saw in §1.3.1 that this need not be the case as there are many different kinds of relations captured by *semantics*. Secondly, there are many ways with which humans can acquire the meanings of new words employing rich inferential mechanisms (e.g., Smith & Yu, 2008; Xu & Tenenbaum, 2007; Yurovsky, Yu & Smith, 2013). While we do not in any way downplay the role of such mechanisms, we conjecture that at an initial stage co-occurrence statistics might be used by children to *bootstrap* their semantic space. Similar ideas were put forward by Hills (2012) who found that the associative structure and contextual diversity in child-directed language facilitated early word learning. Recently, Lazaridou, Marelli & Baroni (2017) have shown that distributional models of semantics augmented by multi-modal models of semantics can induce word meaning in a human-like way. To this end, following Rumelhart & Todd (1993) and Rogers & McClelland (2004), in §7.2 we show that a system learning 'deeper' semantic relations can benefit when using pre-trained co-occurrence vectors in lowering the training time of the system as well as generalising to novel semantic relations.

2.3.2 Recurrent Neural Embeddings

The semantic embeddings presented in §2.3.1 while successful in many tasks can be criticised because they do not take into account linear dependencies inherent in the linguistic structure. Since all the words around a window of size n from the target contribute equally to the prediction, there is no way for the model to 'know' which words followed or preceded the target. Based on recent behavioural, neuroanatomical and computational studies, we have already remarked that we do not consider this to be an issue in the present context. However, for the sake of completeness, we would like to briefly review related methodologies which capture the sequential nature of linguistic phrases. As early as Elman (1990) (also, Elman, 1991) researchers have used *recurrent neural networks* that try to predict each element of a sequence given the element at state t as well as a representation of all the states that preceded. This

method has been quite successful in uncovering how syntactic dependencies might be seen as statistical regularities but could easily scale up to regular sized corpora. The problem with this is that in a corpus with vocabulary size around 1.2×10^5 words the computational bottleneck of computing the Jacobian would, again, be prohibitive. Despite these computational issues, Bengio et al. (2003) has managed to successfully learn a neural language model using a recurrent neural network (similar to Elman). Moreover, Mikolov et al. (2013c) found that when it comes to semantic embeddings, representations as generated by models like the above do not differ from those of a neural language model.

2.4 General considerations

2.4.1 Multiword Expressions

Since the primary unit in distributional semantics models is a space-delimited string, Multiword expressions (MWEs) such as 'SIM card' raise a unique problem. The problem lies in the fact that although the constituent parts of the expression (here 'SIM' and 'card') have representations of their own, there is no representation for the concept of the 'SIM card'. Intuitively, this is wrong as 'SIM card' (the cards used in GSM phones) is something qualitatively different to the 'red card' given in football matches and both of them are not the same to the concept 'card' as in 'After the meeting she left him her card'. A challenge, therefore, for distributional semantics models would be to widen their scope beyond the space-delimited sequence of characters to identify and represent concepts which need more than one word to be expressed.

At the heart of Formal Semantics (Montague, 1970) is the principle that the meaning of the sentence (or a phrase in this instance) can be derived using a rule-governed combination of its constituents. In other words, we can combine these space-delimited atomic units into phrases and sentences using a productive set of rules. Assuming we know what these rules entail, this would provide a helpful framework for which we would be able to derive semantic representations for phrases and sentences automatically. There have been a few attempts to bridge the well-studied field of formal semantics with distributional models (Beltagy, Chau, Boleda, Garrette, Erk & Mooney, 2013; Garrette, Erk & Mooney, 2014) mostly by enriching logical forms with distributional representations.

An alternative would be to identify a set of algebraic operations which would be applied to the semantic representations (i.e. the vectors as derived above) as a proxy for different semantic phenomena such as compositionality, negation, quantification and so on. Various authors have explored this method in detail (Mitchell & Lapata, 2008, 2009, 2010; Polajnar, Rimell & Clark, 2014, 2015) yielding promising results for Compositional Distributional Semantics. In

this strand of research, vector *addition* seems to yield the best results for combining words into phrases. Mikolov, Yih & Zweig (2013d) have independently reached a similar conclusion.

In the present thesis, while we acknowledge the importance of the rule-based approach we use vector *addition* as a way to extract semantic representations for multiword units. We have also considered concatenating the elements of the expressions forming unigrams such as 'simcard'. Considering that the number of multiword units used in the behavioural experiments was small (only 32 MWE) this method could be feasible. However, the frequency of the bigrams (such as 'ID card') was quite low in the corpora yielding uninformed vectors (as shown by similarity tests). Moreover, this method masks the fact that, for example, 'SIM card' and 'ID card' are both cards in some sense, and participants might be aware of this fact during the experiment.

2.4.2 Corpus choice and parameter spaces

All the above models can be trained using any linguistic corpus. For our English simulations, we chose the British National Corpus (British National Corpus (BNC)) as a qualitatively balanced and diverse alternative to the commonly used Usenet and The Touchstone Applied Science Associates corpus (TASA) corpora (see §4.3.2, for other languages). One advantage of the BNC is that it is large enough (ca. 100 million words) but costly operations such as Singular Value Decomposition could still be completed in a very short time.[10] The BNC comprises 4049 marked up texts, and it is a mixture of written texts (comprising of 90% of the corpus) from a variety of domains and a smaller spoken corpus (ca. 10 million words). To make this corpus more DSM-friendly but to incur as minimal information loss as possible we followed the standard practice in the field (Manning & Schutze, 1999), as well as suggestions by Rohde et al. (2006) in performing a series of clean-up steps. Specifically, we performed the following steps

1. removal of XML markup from the BNC files
2. removal of all punctuation marks
3. removal of words over 20 characters in length
4. conversion to lower case
5. automatic spelling correction
6. splitting of hyphenated words

[10]Using the publicly available SVDLIBC library and Intel's Math Kernel Library training an LSA model on the BNC in 200 dimensions took approximately 1h but on 700 dimensions ca. 2.5 days.

We performed steps Item 1 to Item 3 and Item 6 using a set of custom regular expressions. The details of the spelling correction algorithm are given in Rohde et al. (2006), and the implementation was by Peter Norvig.[11] The total number of word types after discarding all words, which appeared five times or less in the corpus was 126097.

We obtain LSA, BEAGLE, COALS semantic vectors using the publicly available S-Space package.[12] As a normalisation procedure (where applicable), we use *term frequency-inverse document frequency*, as described in Table 2.3. Singular Value Decomposition was carried using the SVDLIBC library by Doug Rohde.[13] We also obtained Random Indexing vectors using a custom implementation.[14] Using a custom version of the `word2vec` tool we obtain our neural embeddings. Finally, for the HAL simulations, we use the HiDEx package (Shaoul & Westbury, 2010), which is a configurable implementation of HAL allowing for more control over the original parameters used by Lund & Burgess (1996). The parameter spaces explored for all these models are described in §B.2.

[11] Available at `http://norvig.com/spell-correct.html`.
[12] Available at `https://github.com/fozziethebeat/S-Space/`
[13] Available at `http://tedlab.mit.edu/~dr/SVDLIBC/`
[14] Available at `https://github.com/dimalik/random_indexing`

Chapter 3

Discovering the unconscious representations

3.1 Introduction

The implicit learning phenomena introduced in §1.5, involve recovering the semantically motivated, underlying grammatical system that generated the set of stimuli and making generalisations from it. For now let us assume that simpler explanations based on surface regularities, such as morpho-phonological patterns, cannot explain the generalisation gradients (we test this assumption in §6.2). Developing computational descriptions of implicit learning, therefore, involves using appropriate semantic representations, that capture the effects observed in the behavioural studies. The appropriateness of representations in any context is far from a straightforward issue; take the experiment done by Williams (2005), for example, which we outlined in §1.5. If we construct semantic representations such that they only reflect a single semantic feature [±animacy], then *any* model would exhibit perfect generalisation. Hummel & Holyoak (2003) consider this one of the more severe issues in cognitive modelling, stating that for cognitive modelling to be a 'truly' scientific enterprise there should be a principled way of deriving representations as, otherwise, the modeller can bias the results in their favour.[1]

In §1.3.1 we underlined that not all semantic representations are equivalent as they encode both qualitatively and quantitatively different sorts of information. Word association norms provide sparse information on how words are recalled based on free association experiments. WordNet, on the other hand, provides dense information on how concepts are related based on their hierarchical relations. Not only different models can contain different semantic

[1] In their view, Hummel & Holyoak (2003, p. 247) consider *any* hand-coded representations problematic. While we agree in principle with this assertion, we do not consider representations such as those derived from the McRae norms as hand-coded in the present context. Although they do involve hand-coding, their scale and coverage render them appropriate descriptions of semantic memory.

information, but we also observe this effect within the same model; Landauer & Dumais (1997) notice that in LSA increasing the dimensionality of the vectors can lead to lower fit with the behavioural results (we observe a similar effect in §3.4). However, depending on the needs of the behavioural dataset, such parameters might need tuning to provide better results. As such, they are left free and fitted on a particular dataset.

The objective of the present chapter is to outline and find a principled way of deriving appropriate semantic representations for modelling tasks of semantic implicit learning. On the face of it, this seems like a trivial task; assuming we can construct a computational model for the SIL tasks (see, §5.3.1) we can use representations derived from different methods as input to the models and compare the generalisation patterns of the computational model to the behavioural data. However, there are two problems with this approach; firstly, there is a small amount of semantic implicit learning experiments with different manipulations and cover tasks making it harder to perform such meta-analysis. Secondly, the number of stimuli in these experiments is quite small, which increases the chances of the model overfitting the data.

We find the solution to these issues by making the further assumption that participants do not –at least consciously– activate their semantic knowledge during these tasks. An implication of this hypothesis is that whatever effect we observe *should* be a function of how words are organised by *default* in the mind. Take the *semantic priming* paradigm (Meyer & Schvaneveldt, 1971) as an example. It has been well-established that processing a word can have a facilitative effect on subsequent processing of a semantically related word (e.g., claw → cat) than on an unrelated one (e.g., calendar → cat). As we will explain below, the automaticity of this effect has lead researchers to believe that semantic memory is structured in such a way that by default *evidence* and *trace* are placed closer together. We can now contrast this behaviour to arbitrary tasks of categorisation (e.g., Barsalou, 1983) where participants are asked to find concepts relating to a particular scenario (e.g., 'things to take with you in the case of fire'). While humans can carry out the task giving responses as diverse as 'children', 'dog' and 'blanket', their reaction times as well as the variability in the responses prompt us to think that this is not how concepts are organised in mind.[2] To this end, in this chapter, we focus on deriving the best semantic representations based on tasks of semantic priming in the hope that they will let us model better semantic implicit learning.

[2]This is not to say that ad hoc categories cannot become 'common' ones as frequency plays a major role in their formation (e.g., someone consistently sets their house on fire, so they form the corresponding category) (Barsalou, 1983, p. 244). However, the frequency argument only supports our thesis as within distributional models of semantics words are distributed in space according to the frequency of co-occurrence.

3.2 Semantic Priming

In the present study, the focus lies on accounting for data that reportedly capture the underlying organisation of concepts in semantic memory. We seek to explore the representations participants use by looking at publicly available datasets of *semantic priming* (Hutchison et al., 2013) (in the present Chapter) and *lexical decisions* (in Chapter 4). Semantic priming refers to the facilitative effect of a certain word on processing a semantically related word. For example, processing the word CLAW will facilitate the processing of the word *cat* but processing CALENDAR would not. This facilitation manifests as a faster reaction time to say the word *cat* having seen CLAW (μ = 493.02ms, SE=19.41) compared to when having seen the word CALENDAR (μ = 572.04ms, SE=32.36).[†] Because –apparent– semantic relatedness influences this non-semantic task (reading a word aloud) researchers suggest that this effect is due to the underlying organisation of semantic memory.

Subsequent literature has corroborated these early results (Neely, 1977, 1991, for early reviews) and (Hutchison, 2003a; Lucas, 2000, for more recent meta-analyses) highlighting the importance of semantic priming in understanding the structure of semantic memory. Most of the experiments have focused on finding the kind of relations between the prime and the target that are more likely to yield semantic priming effects. For example, researchers have focused on whether the observed effects are due to semantic overlap or association strength between the prime and the target (Chiarello, Burgess, Richards & Pollock, 1990; Moss, Ostrin, Tyler & Marslen-Wilson, 1995; Williams, 1994), whether mediated relations (e.g., see §2.2.1) can yield semantic priming or whether the effects depend on the input modality (Moss, McCormick & Tyler, 1997; Ostrin & Tyler, 1993; Zwitserlood & Schriefers, 1995). In his meta-analysis, Hutchison (2003a) summarises the main conclusions of these studies highlighting that when care has been taken in choosing the type of association between the prime and the target (e.g., synonyms, antonyms) both semantic and associative relations can yield semantic priming effects, as well as that both cross-modal and mediated priming are also possible.

One problem identified early on (Neely, 1977) was that during the experiments participants notice the relations between the word pairs and form conscious strategies that enable them to predict the upcoming words. This anticipation problem masks the overall priming effect as the faster response to the target would be explained from the participants' predictive strategies rather than from the structure of the semantic memory. There are two main ways of gauging automaticity in such experiments. Firstly, the researcher generates more balanced lists where foils are dispersed alongside critical trials so that the relatedness of some words is not that obvious. Secondly, following Posner & Snyder (1975) and Becker (1980) the researcher can lower the time offset between the prime and the target word. The reasoning behind this decision is that such strategic shifts in attention require time to occur, hence by lowering this

threshold we decrease the chances of participants generating expectancies. While there is no agreed threshold distinguishing between conscious strategies and unconscious processing (Hutchison, Neely & Johnson, 2001), probably because it depends on other variables (e.g., the nature of the task), in the study described below we focus on those trials in the SPP where the onset of the target word was 250ms after the start of the trial.[3]

3.2.1 Prior Work

Very few computational approaches have been attempted to link distributional vectors to semantic priming effects. Lund et al. (1995) correlated the Euclidean distance (Table 2.4) between word vectors generated by HAL (§2.2.2) with previously reported priming effects (Chiarello et al., 1990; Shelton & Martin, 1992), as well as from one novel experiment. While HAL was unable to model associative relations, it was successful at capturing the priming effects of semantic relations. Given the discussion in §2.2 above, we can attribute this preference to the paradigmatic nature of semantic and taxonomic relations (see, Ex. 2.2.3). Moreover, Lund, Burgess & Audet (1996) attempted a more rigorous examination, separating pure semantic, semantic + associative and pure associative relations. They found again that HAL was able to predict similar priming effects for the semantic and semantic + associative relations but not for the purely associative relation. Subsequent studies have focused on whether relations that exert semantic priming are recoverable in DSMs. In one instance, Livesay & Burgess (1998) were unable to show mediated priming effects such as those shown in Balota & Lorch (1986) in HAL. Conversely, Chwilla & Kolk (2002) were able to demonstrate mediated priming effects in LSA. We attribute this dissociation, again, in the models' specifications; the dimensionality reduction step in LSA brings forward such higher order co-occurrences whereas HAL is unable to account for a similar effect.

The first study to systematically compare the predictions of different DSMs with semantic priming effects demonstrated under a variety of conditions comes from Jones et al. (2006). In this study, Jones et al. (2006) train LSA, HAL and BEAGLE models using the TASA corpus and examine the effects predicted by the cosine differences in the word vectors against those obtained in the behavioural experiments. The authors report simulation results from nine different experiments attempting to distinguish between semantic and associative relations. In short, they look at the predictions between semantic and associative relations (Chiarello et al., 1990; Ferrand & New, 2003), more fine-grained semantic relations (e.g., script, collocate or instrument) (Moss et al., 1995; Williams, 1996) and mediated priming (Balota & Lorch,

[3]This number refers to the Stimulus Onset Asynchrony (SOA), that is, the time from the start of the trial until the onset of the target word. The inter-stimulus interval, i.e., the time gap between the prime and the target was 50ms.

1986; de Groot, 1983; McKoon & Ratcliff, 1992; McNamara, 1992; McNamara & Altarriba, 1988). The authors found that BEAGLE was able to make the best predictions overall, as either HAL or LSA underestimated the priming effect of the associatively and semantically related words, respectively. Conversely, the integrative nature of BEAGLE combining two streams of information (paradigmatic and syntagmatic) into a single representation makes it more successful in modelling these relations.

One shortcoming of this study noted by Hutchison, Balota, Cortese & Watson (2008) is that it ignores performance on the item level. That is, Jones et al. (2006) compared *only* the priming effects from the studies with the ones obtained by the computational models. The problem is that considering the size of the datasets, even one pair would be able to skew the results. In their study, Hutchison et al. (2008) using hierarchical regression introduced LSA priming scores (similarity of an unrelated pair - similarity of a related pair) as a factor to the model to explain the variance in standardised scores of priming effects. LSA scores did not enter in the hierarchical regression model as a significant factor, nor did they correlate significantly with priming effects, but LSA managed to predict a priming effect found in the data as related words had significantly higher similarity values than unrelated ones. Although these results might suggest that DSMs are unable to capture the structure of semantic memory, we note that the above study was based solely on the examination of a single LSA model trained on 300 dimensions.

Using the estimates of a single instantiation of a single model might seem all too biasing against LSA as the results are too dependent on the training parameters. Indeed, both Hutchison et al. (2008) and Jones et al. (2006) did not train the LSA models themselves but used the estimates derived by Landauer & Dumais (1997).[4] The appeal of using such pre-trained vectors for this task, on the one hand, ensures that these estimates are 'proven' and not a result of spurious correlations. On the contrary, if the choice of the model or its parameters are dependent on the present task, it follows that we need to tune the model parameters accordingly.

Two more studies appeared recently (Ettinger & Linzen, 2016; Mandera, Keuleers & Brysbaert, 2017) that link DSMs with priming effects. Ettinger & Linzen (2016) train several neural embeddings models on a few configurations of parameters until they maximise the fit between the computational model estimates and the priming effects obtained in the Semantic Priming Project. The authors equate the two tasks (naming and lexical decision) and do not focus on the results of the former as the DSMs do not provide a good fit there. Similar to Hutchison et al. (2008) the DSMs do not seem to improve the fit in comparison to a baseline model containing only word frequency as its covariate. Mandera et al. (2017), on the other

[4] http://lsa.colorado.edu

hand, looked at directly predicting the RTs from the Semantic Priming Project. Using LSA, HAL and neural embedding estimates, they report a higher fit for a baseline model that contained *word frequency*, the *number of orthographic neighbours* and *word length* as its covariates (R^2 = 0.31). Adding DSM estimates improved the model by a small amount (R^2 = 0.33), however, performing rigorous analyses using Bayes factors it turned out that the model was significantly better. The relatively high results obtained by Mandera et al. (2017) might be explained as a result of *overfitting*. They report R^2 values coming from the same set used to find the best parameters for the models. The problem with this approach is that there is no independent test set used for evaluation. We can conjecture from this that is it possible that the models found a 'good' solution regarding the fit but meaningless otherwise.

The aim of this study, therefore, is to find the best representations capturing elements of the semantic memory using VSMs as well as other semantic baselines (described in §3.3.1). Contrary to Ettinger & Linzen (2016) we train several different models on the dataset on many parameter settings. Moreover, we improve on Mandera et al. (2017), again by training more models on more settings as well as following a more rigorous data splitting and model selection procedure. The parameter sets provided by each of these models gives endless possibilities as to what might be the optimal tuning for predicting semantic priming effects. In §3.3.1 we outline a general methodology in the hope of isolating some key parameters (such as sentential and vector size, encoding of syntactic relations) which influence the fit of the models in approximating semantic memory representations.

3.2.2 Impact on models of semantic priming

The issue of whether semantic priming is the result of overlapping semantic representations or simply because of some learnt association between the words has had an impact on computational approaches, too. Take the spreading activation theory (Anderson, 1983; Collins & Loftus, 1975; Quillian, 1967, 1968), for example; within this framework, semantic memory is assumed to consist of concept-nodes which are connected to either other concepts or semantic features via *relational* links. For instance, the concepts *cat* and *animal* are connected via a *labelled* link, whereas *cat* and *fur* are connected via a *has* vertex. According to the spreading activation account, when the node of the prime is activated, the activation is spread along its pathways, with which this node is connected. The short reaction time, therefore, is accounted for by the residual partial activation of the target word. The strength of the priming effect is a function of the number of links 'exiting' the prime node and of the frequency of the connection between the word pairs.

Semantic overlap accounts, conversely, assume that words can be represented as sets of semantic features (McRae et al., 2005) and that the size of the priming effect grows with the size

of the intersection between the two sets. Researchers in this domain (Cree et al., 1999; Plaut, 1995) have used attractor neural networks to model semantic priming effects by observing that the network requires fewer steps to activate the target word given a related prime than an unrelated one. Given these two paradigms, it is not exactly clear how distributional models of semantics could fit in. While DSMs can provide us with a notion of semantic distance between two words they lack concrete semantic features, assumed by both theories, as well as links to a subset of words (they are associated with *all* the words in the corpus).

We have already mentioned attractor networks in §1.3.1. Attractor networks can model *memory retrieval* processes by performing *multivariate linear regression* computing the function $f(\mathbf{x}) = \mathbf{y}$, where $\mathbf{x}$ is some probe, e.g., a phonological representation, and $\mathbf{y}$ is the memory trace (e.g., the semantic representation). Attractor networks model the *memory retrieval* process by computing this function in a *series* of discrete steps. During the first few steps, the input pattern is presented to the network computing the output pattern. In the last few steps, activation continues to propagate to the output layer based on the activations of the previous steps. Concretely, let $\mathbf{x}$ be some input phonological pattern of a word and $\mathbf{y}$ be its distributed semantic representation. Also, let T be the total number of steps required for $\hat{\mathbf{y}}$ (the prediction) to settle into $\mathbf{y}$ (the teacher pattern). The way the network computes $f(\mathbf{x}) = \hat{\mathbf{y}}$ is by minimising the distance of the prediction to the target as we approach T (i.e., $\lim_{t \to T} \cos(\hat{\mathbf{y}}, \mathbf{y}) = 1$).[5]

Given this formulation, it should not be surprising that *semantic priming* naturally arises in these models. Consider that we feed the pattern associated with a particular *prime* to the network, and the output layer has settled on the corresponding semantic representation. At $t = 1$, when presenting the target, $\cos(\hat{\mathbf{y}}, \mathbf{y})$ will be higher if the target is related to the prime. Therefore, it should take fewer steps for the network to settle on a stable pattern in the output layer when presenting a related target than when presenting an unrelated one. Within this framework, semantic priming arises because, given a prime word, the semantic memory activates similar features as to when it needs to parse the target word.

Attractor neural networks can accommodate distributional semantic vectors in their output layer without any alterations in their structure. In this context, the semantic vectors are some memory trace (for more on this idea see also, Kintsch & Mangalath, 2011) that is activated given a *prime*. On the other hand, it is far from trivial to extend *spreading activation theory* to accommodate the above representations. Spreading activation assumes that (a) concepts have a limited number of connections to the other concepts and (b) they are also indirectly connected by mutual semantic features. While we can simulate (a) in this context, in Chapter 4 we discuss ways of extracting semantic neighbourhoods from these representations, it is not clear how

[5] In reality, attractor neural networks use a cost function such as the mean squared error between the predicted and golden outputs. However, it is easy to show that as the error of the network gets smaller, the cosine similarity between the predicted and the golden output is maximised.

we could make (b) work. Distributional models of semantics do not encode concrete features as WordNet does, as there is no single element in their vector that encodes, say, *animacy*, or the type of the relation to the feature. While we can extract such information from the word vectors (Demeester, Rocktäschel & Riedel, 2016; Rogers & McClelland, 2004; Rumelhart & Todd, 1993, as well as §7.2), the output still needs a substantial amount of processing to be compatible with spreading activation theory.

3.3 Method

The methodological details outlined here concern the studies described in §§ 3.4, 4.4 and 4.5.

3.3.1 Model Selection

A drawback of the above models is that they introduce a substantial number of (potentially interacting) hyperparameters that need to be tuned. Hyperparameter tuning is a laborious task, and if done improperly it has the potential of levelling at some local minimum (i.e., a region of the parameter space where just changing one or two values in the parameters would not improve the fit). To illustrate how laborious this task can be, consider that the neural embeddings model we use (`word2vec`) has 8 hyperparameters.[6] Even if each hyperparameter could take either one of two values (i.e., a binary variable) the number of potential parameter sets (and consequently models) would be $2^{10} = 256$.[7] Since 2 is the lower bound of the number of different values a variable can take (most of the parameters are real numbers), the number of possible parameters sets to consider grows prohibitively large for **grid search** (i.e., iterating over the possible parameter sets in a 'loop'). To alleviate most of this problem, we use *Bayesian Optimisation* (Snoek, Larochelle & Adams, 2012) to find the best parameter set, a technique that has been found very useful in finding the parameters of DSMs (e.g., Alikaniotis et al., 2016).

The Bayesian Optimisation algorithm constructs a generative probabilistic model for the parameter space $\mathcal{Z}$ and then exploits this model to find regions in the space which maximise some internal 'fit' function (i.e., $p(y|\mathbf{x}, \mathcal{Z})$, where y is the model 'fit', $\mathbf{x}$ is the parameter vector and $\mathcal{Z}$ the parameter space). One added advantage of Bayesian Optimisation is that the choice for the next step does not rely only on the last evaluation (as in Markov Chain Monte Carlo algorithms) but on *all* the previous steps. The expense of the added computation, however, is mitigated by the speed with which the Bayesian Optimization (BO) converges to the best

[6] Simulation details can be found in §B.2.

[7] To put this number into context, training one DSM in the present setting takes on average one hour, so doing a full pass over all the possible models would take about 42 days.

possible results. Since each DSM takes significant time to run, BO provides a way to explore many parameter sets in the shortest amount of time.

What remains is to define an objective for the Bayesian Optimisation algorithm (that is, what it will maximise). Let us take a step back here and consider the task for the moment. If we are to model *reaction times* or *priming effects* (i.e., Reaction Time $(\text{RT})_{\text{unrelated}}$ - $\text{RT}_{\text{related}}$), we want to find a set of predictor variables, such as *word frequency*, *orthographic length*, etc., that maximise the fit of the regression model to the data. Hutchison et al. (2008) performed hierarchical linear regression entering variables relating to either *prime* or *target* characteristics or some measure of their relation (e.g., *forward association strength* as obtained by the Nelson norms, see §1.3.1). Using this set of variables they were able to derive a set of predictors that best predict semantic priming effects at the item level (the same set of predictors were included in the SPP dataset). A simple proposal could, therefore, be to include those parameters which are predictive of priming effects adding the covariate for the semantic model used. Alternatively, we can add *only* the semantic model covariate in the model and then asses the fit.

Both of the above proposals face a similar problem. Firstly, ignoring the rest of the variables when we construct the objective function we might run into the problem that the best parameters for the DSM might be a function of some other 'simpler' covariate such as word frequency. To see how this might be possible, consider the word vectors in Table 2.1, before any normalisation procedure. The magnitude of each vector is an approximate function of the frequency of the corresponding word. If, therefore, we leave the normalisation procedure as a parameter to fit we might run into the problem of learning a 'hard-to-beat' model only by learning a way to estimate word frequency. Secondly, if we include the variables used by Hutchison et al. (2008) together with the semantic model estimates, we might use a more complicated model than warranted giving rise to misleading (overfitted) R^2 values.

We mitigate the above issues by performing *variable selection* on the dataset before training the semantic models. Concretely, using several psycholinguistic variables to be detailed below (cf. §3.3.2) we perform several *feature selection* procedures to find the best baseline model for the naming task. Once we derive the best possible model, we extract the design matrix (i.e., the predictors) and attach the estimates of the semantic model. Three questions remain unanswered from the above; 1) how do we eliminate variables? 2) what do we mean by 'fit' in the present context? and 3) how do we decide on the best possible model?

Regarding the first question, we avoid using *stepwise* regression, which, although favoured by researchers in the field (Buchanan et al., 2001), can give rise to erroneous *beta* values and are biased to return high R^2 values (Tibshirani, 2011). To this end, we chose to use *lasso* regression which minimises the sum of squares (as in regular regression) constraining the ℓ_1

norm of the *beta* values (the regressor weights) to be lower than some threshold s, solving:

$$\arg\min \sum(\mathbf{y} - \mathbf{A} \cdot \mathbf{X}^{\top})^2 \qquad \text{s.t.} \quad \|\mathbf{A}\|_1 \leq s \tag{3.1}$$

where $\mathbf{y}$ is the dependent variable (the priming effects), $\mathbf{A}$ the design matrix, $\mathbf{X}$ the independent variables, $\|\mathbf{A}\|_1$ the ℓ_1 norm of the covariates, and s a threshold value constraining the magnitude of the weights. The advantage of this approach is that by choosing a low value for s (using cross-validation we determined $s = 0.1$), irrelevant *beta* values are going to be very close to zero effectively being cancelled out.

Regarding the rest of the problems, we can say the following. Firstly, by fit, we mean the coefficient of determination (R^2) given by the model with the selected variables and the DSM estimates on the validation set. Because *lasso* depends on cross-validation, deriving a R^2 value directly from this model is a non-trivial procedure. We find the R^2 value by re-training a standard ordinary least squares regression model with the same parameters as the ones returned by *lasso* and compute the R^2 value there. While this might seem confusing (we train the model to re-train another with the same parameters), Belloni & Chernozhukov (2013) have shown that this is an acceptable procedure for *lasso* and performs "just as well" (i.e., using *lasso* only to zero some coefficients, not as a regression model). We then determine the best possible model by comparing the baseline model and the model with the similarity estimates.

3.3.2 Baselines

Common psycholinguistic measures

A major problem in psycholinguistic research is crafting stimuli sets that vary only in one dimension (Cutler, 1981; Hutchison et al., 2008). Since most psycholinguistic designs involve a factorial design or at least a design where the relevant comparison will be made on separate groups (e.g., high vs. low frequency words, semantically related vs. unrelated) balancing the stimuli such that nuisance variables cannot explain the variance is of utmost importance. For example, the orthographic frequency of the target word influences semantic priming (Becker, 1979). This should not be surprising as one way or the other semantic priming involves lexical access which is affected by frequency. Either not controlling for this effect (by crafting balanced lists) or accounting for this in the analysis might lead to erroneous results on what caused the faster RT in the experiment.

There are many variables related to either the characteristics of the *prime* word or the *target* that might influence the RT. To name a few prime or target length, regularity, consistency, bigram frequency, onset, orthographic neighbourhood, meaningfulness, and concreteness (for a more detailed list, Hutchison et al., 2008) can either facilitate or inhibit lexical access,

exhibiting either a positive or negative correlation to RT. As outlined above, including these factors in the model and testing their significance helps us get a better estimate of the influence of semantic similarity in priming. The factors we test for during the lasso procedure are the ones included in the SPP dataset; (a) bigram frequency, (b) word length, (c) word frequency, (d) number of orthographic neighbours, (e) part-of-speech and (f) nature of the relation between the prime and the target (synonyms, antonyms etc.).

As seen in §A.4 the variables we end up with are the *word length* (`Prime_Length`) and the log-transformed *word frequency* (`Prime_LogSubFreq`) for the prime words and the *word length* (`Target_Length`), the log-transformed *word frequency* (`Target_LogSubFreq`) as well as the *number of orthographic neighbours* (`Target_OrthoN`)[8] for the target words. A linear regression model with these coefficients achieves a baseline $R^2 = 0.323$, which is very close to the variance explained by the baseline model in Mandera et al. (2017) ($R^2 = 0.312$). Given that regression models achieve usually achieve low fit with the data (e.g., Buchanan et al., 2001; Hutchison et al., 2008), this model provides a competitive baseline for our experiments.

WordNet

We also offer two semantic baseline measures where we can compute the semantic similarity between the prime and the target; WordNet and Association Norms. Measuring semantic similarity in WordNet is tantamount to measuring the distance between two nodes in the graph. Concretely, we want to find the path $P = (v_1, v_2, \ldots, v_n) \in V \times V \times \cdots \times V$ from word w_1 to word w_2 by minimising n. While this is a very straightforward way and many efficient path minimisation algorithms exist, it quickly faces the issue pointed out by Resnik (1995) that there is an underlying assumption that the distances between the nodes in the graph are uniform. Consider the representation of *dog* in Fig. 1.4; intuitively, the distance between *dog* and another canine such as *wolf* should be shorter than between the concepts *insectivore* and *pet* (both being sister nodes under `animal.n.01`). However, because both pairs are sister nodes (i.e., they are subsumed by the same ancestor) their path distance is the same.

One set of approaches to overcome this problem take into account the depth in the hierarchy of the concepts in question. The reasoning behind this is that concepts 'deeper' in the taxonomy will are more closely related than those higher up (Sussna, 1993). Approaches, therefore, taken by Leacock & Chodorow (1998) and Wu & Palmer (1994) introduce a normalisation element in the path distance calculation that takes into account the depth of the concepts in the hierarchy. The second set of approaches (Jiang & Conrath, 1997; Lin, 1998; Resnik, 1995) involves incorporating corpus statistics in the similarity function. In short, these methods include information theoretic criteria to capture the probability of encountering w_1 given w_2.

[8]The names in the parentheses indicate the column names in the original datasets.

Comparing the possible methods to capture semantic similarity from WordNet, Budanitsky & Hirst (2006) found the algorithms proposed by Jiang & Conrath (1997) and Leacock & Chodorow (1998) to provide the best fit on two behavioural tasks of similarity ratings. In what follows, we examine *all* of the above metrics in the context of semantic priming.

As we remarked above, WordNet includes *concept-concept* (see §1.3.1) relations instead of *word-word* as the association norms and the DSMs. The problem in the present context is that there is not necessarily an 1 : 1 relationship between the word in the SPP and the concept referred to by the WordNet synset. Take the word *cow*, for example; the first two synsets tagged as *cow* are defined as *"female of domestic cattle: 'moo-cow' is a child's term"* and *"mature female of mammals of which the male is called 'bull'"*. Automatically choosing the intended concept could be done using WordNet's internal sorting mechanism that arranges synsets by frequency. However, this can quickly prove to be problematic as the first synset to appear for the word *table* has the definition '*a set of data arranged in rows and columns*' while the related primes in the SPP for the target word *table* are *chair* and *seat*. Hand-picking the words so as to intuitively match the WordNet definition to the intended use in the SPP is both a laborious and potentially biasing task. We mitigate this problem by implementing the following solution; for any two *related* (in the SPP) words we select the two synsets that maximise the similarity metric while for the unrelated words we choose two synsets at random. The reason for the first decision is rooted in models such as the *spreading activation theory* in that given the activation of the prime word, the target word that is going to be activated is the one that stands closer to the prime. This method was also used by Budanitsky & Hirst (2006) and in the original studies introducing the above similarity metrics. We consider the second decision to be a safer option in the present context than if we had implemented the same solution for the unrelated words as initial results showed this method to be biased by spurious correlations.

Word Association Norms

Apart from WordNet similarity metrics, we use the University of South Florida Free Association Norms to obtain similarity ratings for the *prime-target* pairs. We have already described the procedure to obtain Free Association Norms in §1.3.1 and the relation of those norms to the SPP in §3.4.1. Since the *forward association* strength was used to derive the word pairs, we would expect it to exhibit some negative correlation to the reaction times. That is, the higher the association strength, the lower the reaction time needed to process the target word. In the semantic priming task we are reviewing below, the correlation between the *forward association strength* and the *prime-target* pair is $r = -.09$, $p = 0$. While this estimate might seem small, we do not know yet what effect might be considered high in the present context. Furthermore, in

their analysis Hutchison (2003a) found that the Forward Association Strength (FAS) covariate was significant in their model while the LSA estimate was not.

Together with the *forward association strength* we also obtain *similarities* from the association norm representations. Given a cue word w, its associates $\mathcal{A}$ and an associates vocabulary $\mathcal{V}$, we form a sparse vector $\mathbf{w} \in \mathbb{R}^{|\mathcal{V}|}$ such that all its elements are zero except for those that exist in its associates set $\forall i \in \mathcal{V},\ i \in \mathcal{A} \Rightarrow w_i = 1$. Alternatively, the value of the element can be a function of the relationship between the two words. We also explore the use of *forward association strength* as an alternative function Because of the sparsity of the vector, we perform dimensionality reduction via SVD (§2.2.1) to obtain lower dimensionality representations which, hopefully, contain richer information. For the experiments reported here we derive vectors of size 10, 30, 50, 100, 200, and 300. Due to the small number of possible hyperparameter combinations for both WordNet and the Association Norms, we do not use the Bayesian Optimiser on these model as we can derive the estimates directly.

3.4 Study 1: Priming Effects in the Semantic Priming Project

3.4.1 Dataset

We obtain *semantic priming* data using the SPP, a publicly available dataset of reaction times.[9] The SPP contains data from two semantic priming tasks; a *naming* and a *lexical decision*. Since the magnitude of *lexical decision* tasks can be largely task-specific (Neely, 1991), and have been found to strongly correlate with the semantic neighbourhood density (Mirman & Magnuson, 2008) of each word, we explore them further in Chapter 4. For this study, therefore, we focus on the reaction times from the naming task. Moreover, the SPP contains data on two SOA conditions; a short (200ms after the prime) and a long one (1200ms) gathered over two sessions per participant. Considering the discussion above, we mainly focus on the short condition as this is less susceptible to conscious strategies used by participants during the experiment.

The 3322 unique related prime-target pairs were obtained in the SPP using the Nelson norms (Nelson et al., 2004) (and §1.3.1). While the number of word pairs was 3322, only 1661 target words were used. Every target word received either its *first-associate*, as derived from the Nelson norms, as its prime or *any* other associate from the same list. Unrelated targets were, then, generated from randomly re-pairing the prime-target pairs with the additional constraint that the target is not associated to the prime in the word association norms. Hutchison et al. (2013) outline in detail the constraints used to generate the pairs.

[9] `http://spp.montana.edu/`

We extract our dependent variable by selecting all the trials in the 200ms SOA condition dropping incorrect responses. Subsequently, we drop all reaction times longer than 3 standard deviations above the mean. Finally, we z-transform each participant's reaction times by session then averaging by items. We z-transform the reaction times to counter a common problem in reaction time data in that participants differ on their 'baseline' speed. That is, although some participants are faster or slower than others they might show differences between conditions (unrelated and related). A common solution is to transform the reaction times for each participant to their standardised z-scores. This way, all participants share a common 'baseline' around 0, and from that point, we can average by-item. Although this is an acceptable practice and adopted by the SPP, researchers in the field (Baayen, Davidson & Bates, 2008; Baayen & Milin, 2010) have argued against it and in favour of a more theoretically sound approach which is to treat participants as *random effects* and factors relating to the nature of the items (e.g., frequency or semantic similarity) as *fixed-effects*. Using different intercepts for each participant we adjust their baseline RT, achieving a similar result. While we find this approach valuable as it overcomes issues associated with traditional F_1 and F_2 analyses (Baayen et al., 2008) it is practically difficult to run on the present dataset because of its size. While running a linear regression model is computationally cheap taking advantage of tested algorithms performing QR factorisation, linear mixed effects models with multiple random effects are harder to estimate as there are no closed form solutions and approximations can be very expensive as the size of the dataset grows. Further to that, the problems associated with selecting the random effects would cause problems in our model selection procedure outlined in §3.3.1.

3.4.2 Splitting the dataset

Using the same dataset for (1) finding the best parameters for the DSMs, (2) training the regression models and (3) assessing the fit of the models could be susceptible to problems related to *overfitting*. That is, the model we would end up with would have been trained to predict a particular set reaction times best and not to provide a more *unbiased* estimator of those reaction times. Considering the large size of the dataset, we are allowed to split the dataset into different sets which we would use to find the parameters for the models, *train* the models and examine their generalisability on a novel set. Using standard machine learning methodology, we split the dataset into *three* parts:

1. **Training** set (64% of the original data); this is used to train the linear regression models and find the beta values for each of the parameters.

2. **Validation** set (16% of the original data); this is used to find the best model parameters. That is, we train the linear regression models with the goal of maximising the fit to this set.

3. **Testing** set (20% of the original data); this is the set we use to assess the fit of the regression models. Neither the regression beta values nor the DSM parameters have *seen* this set before, ensuring unbiased estimation of the fit.

The above data splitting procedure further provides unbiased results but requires extra care so that the resulting sets are not themselves biased. That is, significant differences between sets regarding some variable of interest might return erroneous estimates as over- or under-estimation of the fit. To this end, we follow a customised randomisation and splitting procedure which ensures that there are no significant differences between some key variables between the sets. More specifically, we guarantee that the different semantic relations between *primes* and *targets* are proportionately represented as well as that the word frequencies are 'roughly' similar between sets. We validate the usefulness of this procedure by running separate linear regression models testing for interactions between `set` (i.e., the data split) and the target variable (i.e., the reaction times). None of the interactions came up significant rendering this data splitting method successful (more details on the procedure and the tests can be found in §A.4).

3.4.3 Results and discussion

We split our results into two sections; firstly we look at the results of the Bayesian Optimiser to determine which model could better predict the RTs for the testing set. We then proceed to look at whether the models are able to capture the priming effects (i.e., $\text{RT}_{\text{unrelated}}$ - $\text{RT}_{\text{related}}$). We also perform a more rigorous analysis, looking at the by-relation priming effects.

Determining the best fit

Table 3.1 shows the overall results for the DSMs simulations on the testing and validation sets. Recall that the difference between the two sets is that the parameters of the model were tuned with respect to the validation set, whereas the testing set provides novel data for the model. Aside from the R^2 fit values and the difference in fit between the two models (ΔR^2), we compare each model against the baseline by testing whether the reduction in the residual sum of squares is statistically significant or not. Furthermore, we offer the Bayes Factors between the model containing the similarity covariate and the baseline (see below). Finally, Table 3.3 shows the parameter sets for the best performing models.

Table 3.1 Summary of the best performing models on the validation and testing sets of the SPP. The predictor variables for each of these models were the same as the baseline model (see text) with the addition of the similarity estimates of the corresponding DSM. B_{m0} shows the Bayes Factor with default mixture-of-variance priors comparing the model that includes the semantic similarity covariate against the baseline. The significance levels refer to the difference between these models and the baseline (see text).

Model	Validation set			Testing set		
	R^2	ΔR^2	B_{m0}	R^2	ΔR^2	B_{m0}
AN_U	0.3	$+0.01^{***}$	2.33×10^{2}	0.33	$+0.01^{***}$	7.07×10^{2}
AN_W	0.29	$< 0.01^{*}$	8.88×10^{-1}	0.33	$+0.01^{***}$	8.36×10^{2}
BEAGLE	0.31	$+0.01^{***}$	1.04×10^{4}	0.34	$+0.01^{***}$	1.73×10^{4}
COALS	0.28	$< 0.01^{**}$	5.82×10^{0}	0.32	$+0.02^{***}$	1.38×10^{4}
HAL	0.32	$+0.03^{***}$	7.62×10^{6}	0.33	$+0.01^{***}$	4.19×10^{4}
LSA	0.31	$+0.01^{***}$	2.64×10^{2}	0.37	$+0.03^{***}$	1.67×10^{8}
Neural Embeddings	0.36	$+0.02^{***}$	3.77×10^{5}	0.37	$+0.02^{***}$	5.06×10^{7}
Random Indexing	0.31	$+0.01^{***}$	2.15×10^{4}	0.34	$+0.01^{***}$	3.96×10^{5}
WordNet	0.29	*n.s*	2.34×10^{1}	0.29	*n.s.*	2.32×10^{-1}

Significance levels: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Note: AN_U = Association Norms (Unweighted); AN_W = Association Norms (Weighted); BEAGLE = Bound Encoding of the Aggregate Language Environment; COALS = Correlated Occurence Analogue to Lexical Semantics; HAL = Hyperspace Analog to Language; LSA = Latent Semantic Analysis; WordNet estimates were obtained using the similarity measure proposed by Resnik (1995) (see §3.3.2).

Interestingly, the models performed consistently better on the testing set than on the validation, despite being trained to maximise their fit to the latter. This points to the direction that the testing set provided an 'easier' dataset for the models, probably because its values were closer to the training set (where the model coefficients were computed) than to the validation. The results overall present the encouraging view that DSMs are able to capture unique variance in tasks of semantic priming, even when we partial out the effect of other well-known psycholinguistic variables and we test on a novel dataset. The best performing models were the Neural Embeddings, LSA and HAL with the Neural Embeddings being more consistent across splits. These are followed by Random Indexing and BEAGLE and the unweighted Association Norms. Worst performing models were the weighted Association Norms, COALS and WordNet. Apart from LSA, the other models based on associative relations (COALS and the association norms) fared worse than the *paradigmatic* models. We find two plausible

Table 3.2 Pearson correlation coefficients between the similarity measure of each DSM and the priming effects in the naming task of the SPP.

Model	Validation Set	Test Set
AN_U	-0.07^{*}	$-0.04^{\dagger}$
AN_W	-0.09^{**}	-0.07^{***}
BEAGLE	-0.13^{***}	-0.1^{***}
COALS	-0.07^{*}	-0.09^{**}
HAL	-0.11^{***}	-0.1^{***}
LSA	-0.09^{**}	-0.08^{**}
Neural Embeddings	-0.19^{***}	-0.21^{***}
Random Indexing	-0.11^{***}	-0.1^{***}
WordNet	0	0

Significance levels: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

explanations for these results; firstly, due to the nature of the relations mostly eliciting semantic priming effects (see Hutchison, 2003a), paradigmatic models where able to perform better. Alternatively, the unequal distribution of semantic relations in the SPP (see, Fig. 3.1) biases the results towards models that can capture specific semantic relations better. We revisit this point in later in this section and in §3.5 where we look at other datasets controlling the kind of relations include.

With the exception of WordNet, all the models provided a significant improvement over the baseline model in both the validation and the testing sets. Becuase of this performance, we base our discussion of the results on the Bayes Factors also reported in Table 4.2. The unweighted Association Norms yielded a small improvement in both the validation ($F(1, 1015) = 15.95, p < 0.001$) and the testing ($F(1, 1276) = 19.05, p < 0.001$) sets, whereas the weighted Associated Norms even smaller in the validation ($F(1, 1015) = 4.85, p < 0.05$). Interestingly, BEAGLE also had a small effect on either set (validation: $F(1, 1043) = 25.15, p < 0.001$, test: $F(1, 1309) = 25.07, p < 0.001$), despite being able to capture the priming effects in (Jones et al., 2006). This supports the concerns raised by (Hutchison et al., 2008) that including item level estimates is important in modelling semantic priming. Followingly, COALS, another DSM based on syntagmatic relations, yielded a very small improvement in the validation set ($F(1, 765) = 8.48, p < 0.01$) and performed comparably well to the rest in the testing ($F(1, 948) = 24.41, p < 0.001$). So far, the results give the quite clear picture that syntagmatic models perform consistently worse than paradigmatic in the present task. This is somewhat complicated by the performance on LSA on the testing set

($F(1, 1261) = 53.29, p < 0.001$) (validation: $F(1, 1015) = 17.31, p < 0.001$). However, given that the test set was in general 'easier' for all the models, we consider this performance more of an outlier. Turning to paradigmatic models, HAL and Random Indexing performed comparably well with HAL being somewhat better on the validation set ($F(1, 910) = 37.55, p < 0.001$) than on the testing ($F(1, 1121) = 26.18, p < 0.001$). Finally, however, the neural embeddings were consistently high in both sets (validation: $F(1, 1058) = 31.21, p < 0.001$, testing: $F(1, 1277) = 41.48, p < 0.001$) and seem to provide the best alternative in modelling semantic priming effects.[10] Table 3.2 shows the correlation coefficients between the priming effects and the similarity estimates for each of the models. While some discrepancies are present due to the presence of other variables in the linear regression, the neural embeddings remain the highest scoring model.

None of the models containing any of the WordNet variables (that is, the different similarity metrics) was able to surpass the baseline model, nor any similarity metric was a significant predictor in any of the sets. Pearson correlation coefficients between the priming effects and the Resnik similarity metric (Resnik, 1995), which was found to yield the best results, were also statistically insignificant ($r(1084) = -0.03, p = 0.25$) although in the correct direction (i.e., predicting facilitation instead of inhibition). However, an interesting point that we revisit later in the discussion is that the correlation between the WordNet similarity metric and the reaction time becomes significant in the 1200ms part of the SPP ($r(1081) = -0.09, p = 0.001$) suggesting that WordNet might favour more conscious strategies (developed in the 1200ms SOA condition) than automaticity.

Both the weighted and the unweighted versions of the association norms entered as significant predictors in their respective linear models. We further examine whether or not we need to weigh the elements of each vector by the probability of a word appearing as a target to a given cue. That is, using the *forward association strength* instead of a binary variable. Firstly, as a simple metric, we compare the AIC values of the two models where the unweighted model fares better (unweighted: −4.56, weighted: 6.48). However, since there is commonly disagreement on whether AIC should be used to compare non-nested models (as in the present case), we also use a Cox test for comparing non-nested models (Davidson & MacKinnon, 1981). The idea behind the test is quite simple; if one model (e.g., the unweighted) contains the correct set of regressors, then adding the fitted regressors of the other model should not increase the explanatory value. The addition of the unweighted fitted values to the weighted

[10]Some discrepancies in the degrees of freedom are due to either errors in the dataset (missing values in some conditions) or because some similarity metrics or models could not compute similarity for a word. This latter effect is observed when the word is either too rare or too common as some pre-processing steps in the models require dropping those words.

Table 3.3 Best parameter sets for each of the DSMs reported in Table 4.2. The parameters for BEAGLE, COALS, HAL, LSA, neural embeddings and random indexing were found using the Bayesian Optimiser. For the association norms we could directly explore the fit of different sizes after the dimensionality reduction step. Finally, for WordNet we compared directly six different similarity metrics described in §3.3.2. Each model was fitted independently and the baseline was nested in each of them. See text for more information on the parameters of each model.

Model		Parameter	Value
Association Norms	Unweighted	Dimensionality	300
	Weighted	Dimensionality	100
BEAGLE		Dimensionality	256
		Semantic Type	Context
COALS		Dimensionality reduction	No
		Original Dimensionality	14000
HAL		Dimensionality	14000
		Window size	4[1]
LSA		Dimensionality	300
Neural Embeddings		Dimensionality	300
		Window size	6
Random Indexing		Dimensionality	256
		Window size	3
WordNet		Similarity metric	Resnik Similarity

[1] Window size of 4 means four words before and four after the target.
Note: Since none of the WordNet models provided a significant improvement over the baseline, the similarity parameter reported here is the one that minimised the Akaike Information Criterion (AIC) value on the validation set.

model, resulted in a better model ($z = -6.02$, $p < 0.001$) indicating that there were statistically significant differences between the two methods on this task.

The Bayes Factor analysis (Rouder & Morey, 2012) shown in Table 3.1 indicates that despite yielding significant reduction in the residual sum of squares, several models are not on the same scale as others. Since the Bayes Factors were calculated as in (3.2) they denote the probability of the data under the model containing the similarity covariate ($\mathcal{M}_1$) relative to

the baseline ($\mathcal{M}_0$).

$$B_{m0} = \frac{p(\mathbf{y}|\mathcal{M}_1)}{p(\mathbf{y}|\mathcal{M}_0)} \tag{3.2}$$

Under this definition, the probability of the data under the model with the unweighted association norms estimates is 233 times higher than the baseline on the validation set and 836 times on the testing. Comparatively, $\mathcal{M}_{\text{HAL}}$, the model containing the HAL estimates, is about 7.5 million times more probable on this data than the baseline. It is evident, therefore, that together with WordNet (which was not a significant predictor in the traditional analysis), the Association Norm estimates (both weighted and unweighted) are on a different scale than the DSMs (in the order of a few thousand).

Interestingly, the LSA model was also a significant predictor, contrary to Hutchison et al. (2008). We attribute this dissociation to the fact that here we fit the LSA vectors to the task, instead of using pre-trained vectors. To further examine the relation between dimensionality in LSA and model fit, we construct a linear regression model entering the validation score as the dependent variable and the different vector sizes explored by the Bayesian Optimiser as the independent. Vector size entered as a marginally significant predictor of the score in the validation set ($\beta = -0.63$, $SE = 0.31$, $p = 0.09$). The directionality of the sign in the beta coefficient indicates a negative correlation[11] between dimensionality and score, where higher dimensionality results in lower scores. This result is similar to Landauer & Dumais (1997) who found that increasing the dimensionality of the representations results into less agreement to behavioural measures. Similar to these authors, we interpret this finding by conjecturing that the dimensionality reduction introduces more noise to the representations instead of enriching them.

Regarding the BEAGLE model, we used only the *contextual* representation as the source of semantic representations. This is done because we leave the nature of the representation (*ordering*, *contextual*, *composite*) as a free parameter to the Bayesian Optimiser to select. We recognise that this might be problematic as BEAGLE was designed to capture both streams of information, but considering the nature of the task we might have been biasing against BEAGLE if we forced to it include both *ordering* and *contextual* information. As it happens, the Bayesian Optimiser selected only the *contextual* part of BEAGLE as the best alternative (at least on the present dataset).

The above results suggest that when predicting reaction times word to word associations (as computed by the DSMs and the association norms) are more important than concept

[11]Since we include only one parameter in the model, the standardised beta coefficient of the predictor (reported here) is equal to the Pearson product-moment correlation coefficient.

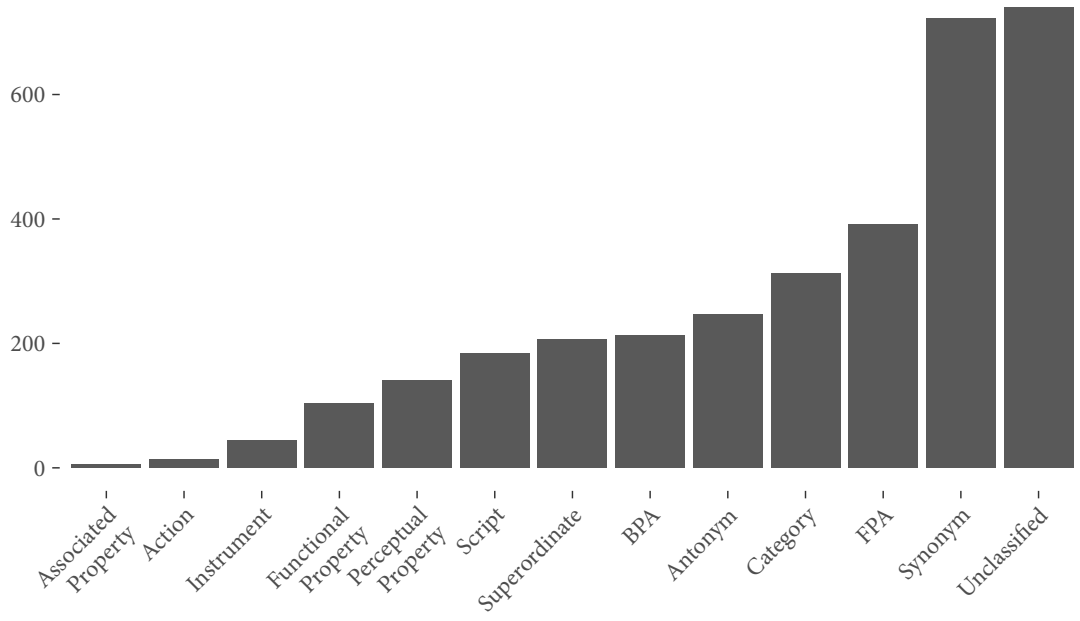


Figure 3.1 Distribution of the semantic relations included in the SPP ordered by number of pairs. The total number of prime-target pairs in the SPP was 3322 (see §3.4.1). *Note:* Backward phrasal associate (BPA) and Forward Phrasal Associate (FPA) stand for *backward* and *forward* phrasal associates, respectively. We omit 'Associated Property' and 'Action' from further by-relation analyses as there are fewer than five pairs related this way.

to concept relations. We move now to look at how these models capture specific semantic relations. While the SPP provides a helpful dataset of human reaction times, running machine learning algorithms to uncover semantic representations might lead to various kinds of spurious errors. For example, Table 3.4 and Fig. 3.1 show the different semantic relations included in the SPP along with the average priming effect on the behavioural task. We see that different models can capture different semantic relations. If the fitting procedure has managed to maximise the overall fit of a model solely by looking at one specific relation, then it does not mean that the model is an accurate approximation of the semantic memory, just of one aspect of it.

Looking at the semantic relations

In order to examine the fit in specific semantic relations, we look at the entire dataset omitting the validation set along with any relation with five pairs or less. Table 3.5 displays the average human priming effects and model predictions aggregated by semantic relation for each of

Table 3.4 Examples of the semantic relations contained in the SPP (see also, Fig. 3.1).

Relation	Example
Action	*scrub–dishes*
Antonym	*day–night*
Associated Property	*dark–cold*
Backward Phrasal Associate	*boy–baby*
Category	*table–chair*
Forward Phrasal Associate	*baby–boy*
Functional Property	*broom–sweep*
Instrument	*broom–floor*
Perceptual Property	*canary–yellow*
Script	*restaurant–wine*
Superordinate	*dog–animal*
Synonym	*afraid–scared*
Unclassified	*mouse–cheese*

the models. We calculate each cell of the table by subtracting the estimate for the related condition from the corresponding estimate for the unrelated condition. For the human results, we do so by subtracting the reaction times (i.e., $RT_{UNRELATED}$ - $RT_{RELATED}$), where a *positive* value denotes facilitation. In the case of the computational models, we record the cosine difference between the two conditions in which case a *negative* value denotes facilitation. Apart from the associative relations, all the other models manage to capture some general characteristics of the human PES, such as that the *category* and *antonym* relation produce the highest PES, while *perceptual property* and *instrument* the lowest. However, only the neural embeddings manage to predict the inhibitory effect in the *instrument* relation, while they overestimate the PES of *category* and *antonym* relations. It might be surprising though that LSA also predicts a highly facilitative effect for the *category* and *antonym* relations, which are considered towards the paradigmatic end of the relational spectrum. We interpret this inconsistency as supporting Hutchison (2003a) and Moss et al. (1995), who argue that none of these relations is clear-cut either *associative* or *semantic* but that they are a collection of different relations.

Turning to specific correlations by relation, Table 3.6 shows that the predictive neural network managed to reach significance in most relations, followed by Random Indexing (RI) and BEAGLE, while LSA, HAL and COALS perform rather poorly. From the Association Norms only the unweighted version significantly correlates with the priming effects. However, as in the case of the two significant WordNet correlations, for three of the relations the models predict an inhibitory effect as denoted from the sign of the correlation. Since the effects

Table 3.5 Average human priming effects and model predictions aggregated by semantic relation. We calculate each cell by subtracting the related condition from the unrelated one. For the human results, this is done by subtracting the reaction times, where a positive value denotes facilitation, whereas for the computational models is the cosine difference between the two conditions (hence, a negative value denotes facilitation).

Relation[1]	H	AN_U	AN_W	BEAGLE	COALS	HAL	LSA	NE	RI
Antonym	+13	−0.35	−0.50	−0.14	−0.24	−0.20	−0.28	−0.19	−0.09
BPA	+ 9	−0.23	−0.24	−0.06	−0.1	−0.06	−0.18	−0.09	−0.06
Category	+14	−0.46	−0.54	−0.12	−0.22	−0.11	−0.26	−0.19	−0.09
FPA	+13	−0.19	−0.23	−0.05	−0.11	−0.09	−0.20	−0.11	−0.04
Functional Property	+ 8	−0.33	−0.43	−0.01	−0.07	−0.04	−0.18	−0.08	−0.02
Instrument	− 1	−0.31	−0.37	−0.01	−0.07	−0.01	−0.20	+0.02	−0.01
Perceptual Property	0	−0.24	−0.24	0	−0.08	0	−0.17	0	−0.01
Script	+ 9	−0.33	−0.36	−0.07	−0.10	−0.03	−0.23	−0.1	−0.05
Superordinate	+ 4	−0.36	−0.39	−0.06	−0.13	−0.06	−0.20	−0.06	−0.05
Synonym	+ 9	−0.41	−0.45	−0.11	−0.14	−0.07	−0.16	−0.11	−0.08
Overall	+ 9	−0.3	−0.32	−0.09	−0.14	−0.08	−0.20	−0.12	−0.06

Note: The human estimates were computed on the raw reaction times instead of the z-scores. BPA = Backward Phrasal Associate; FPA = Forward Phrasal Associate; Neural Embeddings (NE) = Neural Embeddings.

[1] Relations with less than five elements (Action and Associated Property) were omitted from the analyses.

are rather small, we interpret these inconsistencies as spurious correlations in the dataset. The results for some of the DSMs present an interesting inconsistency with our previous results, where HAL and LSA outperformed BEAGLE and Random Indexing in the overall results. However, given how we obtain these estimates, such inconsistencies are expected. Recall that the Bayesian Optimiser was given the reaction times $\mathbf{y}$, a design matrix $\mathbf{X}$, a model class $\mathcal{M}$, a parameter space $\mathcal{Z}$, and sought to maximise $P(\mathbf{y}|\mathbf{X}, \mathcal{M}, \mathcal{Z})$. We proceed to explain this inconsistency as the optimiser finding a configuration of parameters $z \subset \mathcal{Z}$ that achieves high fit with part of the dataset while performing poorly on the rest. Table 3.6 shows that this part of the dataset would be the *instrument* (despite its small size it achieves good fit) and *antonym* relations. It is not necessarily the case that this performance reflects the fact that the Bayesian Optimiser was 'stuck' on a local minimum. Given the unequal distribution of relations in the dataset and the models' idiosyncrasies, the best estimate might not be one that models each

relation but one that is skewed towards one portion of the data. We take this as an indication that no matter how sophisticated some machine learning techniques can be, thay will probably fail without knowledge of the data.

Two further remarks need to be made that can contribute to our understanding of the low scores in COALS, HAL and LSA as well as the inconsistency in the LSA scores in Table 3.5. Firstly, since the best parameters were determined on the reaction times and not the priming effects, there is an equal amount of unrelated pairs that are being modelled. In other words, we do not only determine which pairs should yield priming effects but also which pairs *should not*. The correlations between the reaction time z-scores and the similarity estimates for the unrelated condition were significant (or predicting the correct sign) for fewer models (the neural embeddings, HAL, LSA and slightly less for Random Indexing). A second remark is that a word pair can fit several different categories at the same time. For example, the words *attic* and *basement* can fit both the *antonym* and the *category* relation. While in the SPP there has been an effort to list more than one relation between word pairs it seems plausible that an exhaustive enumeration is not possible. The above remarks suggest that the specific inconsistencies are not results of spurious correlations but rather a lack of correspondence between relation labels in the models and the SPP.

3.5 Study 2: Mediated Semantic Priming

Despite the success of the above models in predicting semantic priming effects directly, the dataset provided was not balanced in the semantic relations it contains. The problem we identified from Fig. 3.1 is that some of the models might be successful in capturing one or two relations thus achieving good results (see also, Table 3.5) without capturing established semantic relations (as found, for example, in Hutchison, 2003a). Furthermore, the SPP does not contain stimuli that would yield mediating priming effects (e.g., *lion* primes *stripes* through its association to the word *tiger*), a well-documented type of semantic priming (Balota & Lorch, 1986; de Groot, 1983; McNamara & Altarriba, 1988). We explore the performance of the best performing above models on three different datasets of mediated priming examining their predictions against the priming effects obtained in the behavioural studies.

3.5.1 Method

In these simulations, we use the best performing models of the previous study and apply them to novel datasets without further tuning of the parameters. We omit WordNet and the Association Norms from further experiments because of their low performance. Similar to the

Table 3.6 Pearson correlation coefficients between the priming effects in the SPP and the model predictions aggregated by relation. Statistically insignificant correlations ($p > 0.05$) were omitted from the table.

	AN_U	BEAGLE	COALS	HAL	LSA	NE	RI	WordNet
Action					-0.78^{**}	-0.67^{*}		
Antonym		-0.15^{*}				-0.25^{***}	-0.14^{*}	0.17^{*}
BPA	-0.14^{*}							
Category	0.18^{**}	-0.13^{*}		-0.15^{*}		-0.2^{***}	-0.11^{*}	
Instrument		-0.38^{*}					-0.36^{*}	
Perceptual Property		-0.2^{*}	-0.31^{**}		-0.2^{*}	-0.22^{*}	-0.2^{*}	
Script				-0.17^{*}			-0.19^{*}	
Superordinate	0.15^{*}					-0.14^{*}		
Synonym	0.08^{*}				0.09^{*}	-0.08^{*}		0.1^{*}

Significance levels: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 3.7 Studies examined in Jones et al. (2006) and subset of studies included here with the corresponding semantic manipulation.

Study	Semantic manipulation
Balota & Lorch (1986)	First-order mediated priming
McNamara (1992)	Second-order mediated priming
McKoon & Ratcliff (1992)	Compound cues

above, we find the priming effects by subtracting the cosine similarity of the related pair from the unrelated pair. Since the included studies only report the related prime-target pairs, we construct unrelated pairs by shuffling the list of primes in each condition with the constraint that the same word cannot be a prime in the related and unrelated conditions.

3.5.2 Dataset

We examine the performance of our models on three studies (Balota & Lorch, 1986; McKoon & Ratcliff, 1992; McNamara, 1992) of mediated priming included in Jones et al. (2006) (Table 3.7). While we could attempt a more rigorous examination looking at the reported priming effects on balanced datasets testing for specific relations, preliminary results gathered from the datasets looked by Jones et al. (2006) show that the priming effects predicted in Table 3.5 hold across studies.

3.5.3 Results and discussion

Balota & Lorch (1986)

The first study on mediated priming we look at was reported in Balota & Lorch (1986) and compares the priming effects obtained between directly related pairs (e.g., *tiger* → *stripes*) and pairs related through a mediating concept (e.g., *lion* → *tiger* → *stripes*) using a set of unrelated primes as their baseline. Crucially, the authors use the same target words in both conditions changing only the list of primes accordingly. The list of unrelated primes is constructed by shuffling the prime words in each condition. Of interest to the present study are the reported results from the naming task ("Pronunciation Experiments", in their paper) performed on two SOA conditions (250 and 500ms). The authors report facilitation in both the related and mediated conditions (as compared to the unrelated baseline) but a difference between them (Relation > Mediated) only on the 250ms SOA condition. Here we report the mean RTs in the

Table 3.8 Data from Balota & Lorch (1986) and corresponding model predictions. The human data show RTs in each condition, whereas for the computational models we report the average cosine similarity between the *prime* and the *target* (the numbers in parentheses denote standard error of the mean).

Model	Related	Mediated	Unrelated
Human	549	558	575
BEAGLE	0.79 (0.02)	0.8 (0.02)	0.78 (0.02)
COALS	0.57 (0.03)***	0.5 (0.03)***	0.42 (0.03)
HAL	0.69 (0.03)	0.69 (0.03)	0.69 (0.03)
LSA	0.30 (0.03)***	0.19 (0.02)*	0.14 (0.02)
Neural Embeddings	0.35 (0.02)***	0.24 (0.01)***	0.19 (0.01)
Random Indexing	0.88 (0.01)	0.88 (0.01)	0.87 (0.01)

Significance levels: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$
Note: Significance testing was performed comparing the two groups against the baseline (using paired t-tests). See text for more comparisons.

naming task collapsing across SOAs. This latter decision was made to avoid extra clutter in the presentation since the direction of the effect remains unchanged.

The neural embeddings predict facilitation in both the related ($t(47) = 8.38, p \approx 0.001, d = 1.22$) and the mediated ($t(47) = 5.04, p \approx 0.001, d = 0.74$) conditions compared to the unrelated controls. Furthermore, similar to the behavioural results there was a statistically significant difference between the two conditions ($t(77.2) = 4.57, p \approx 0, d = 0.66$). Random Indexing vectors predict a null effect in both the mediated ($t(47) = 0.81, p = 0.42, d = 0.17$) and the related ($t(47) = 0.1, p = 0.92, d = 0.02$) conditions. Considering that these vectors contain information on the paradigmatic relations, this is not surprising (cf. Livesay & Burgess, 1998). Contrary to previous results, BEAGLE also does not predict facilitation in any of the conditions (Related: $t(47) = 0.24, p < 0.81, d = 0.03$, mediated: $t(47) = 1.16, p < 0.25, d = 0.15$) compared to the unrelated controls. Furthermore, the two conditions did not differ significantly from one another ($t(47) = -0.78, p > 0.05, d = 0.08$). COALS predicts facilitation in both the related ($t(40) = 4.39, p < 0.001, d = 0.69$) and the mediated ($t(39) = 3.58, p = 0.001, d = 0.57$) conditions compared to the unrelated controls. Furthermore, the two conditions did not differ significantly from one another ($t(78.78) = 0.46, p > 0.05, d = 0.07$). HAL vectors were unable to predict facilitation in either the related ($t(37) = 0, p \approx 1$) or the mediated ($t(38) = 0.01, p \approx 1$) conditions compared to the unrelated controls. Furthermore, the two conditions did not differ significantly from one another ($t(72.57) =$

$-0.01, p = 0.99, d = 0$). LSA predicts facilitation in both the related ($t(47) = 4.8, p \approx 0, d = 0.7$) and the mediated ($t(47) = 2.34, p < 0.05, d = 0.34$) conditions compared to the unrelated controls. Furthermore, the two conditions were marginally significantly different from one another ($t(93.92) = 1.82, p = 0.07, d = 0.26$).

In sum, only the two syntagmatic models and the neural embeddings were able to predict facilitative effects in either condition. However, only the neural embeddings were able to predict a significant difference between the two conditions in the short SOA condition of the behavioural results (250ms). The reported effect sizes in each comparison complement this picture as they all indicate a higher effect in the Related condition, followed by a smaller one in the Mediated and an even smaller in the comparison between the two conditions.

McNamara (1992)

McNamara (1992) explores whether long-distance mediation between the prime and the target still exerts semantic priming effects. An example of this would be *mane* → *stripes* through the intervention of *lion* and *tiger*. *Mane* and *lion* are linked by a meronymic relation, while *lion* and *tiger* by a semantic, and, again, *tiger* and *stripes* by another meronymic. Within the spreading activation (Collins & Loftus, 1975) theory discussed above (§3.2.2), when *mane* is activated, it starts a ripple activation propagating to the words that is connected which, in turn, activate, albeit with a smaller magnitude, the words to which they are connected. Despite the activations being subtle at this point, they would still be more activated than unrelated primes. McNamara (1992) reports a 10ms facilitation effect from a *lexical decision* task for the mediated primes (597ms) compared to the unrelated condition (607ms).

The results obtained in this study are very close to the ones from Balota & Lorch (1986). The neural embeddings predict facilitation for the long-distance mediated pairs ($M_{\text{Related}} = 0.18, SE = 0.02, M_{\text{Unrelated}} = 0.14, SE = 0.01, t(39) = 2.65, p = 0.01, d = 0.42$). BEAGLE, on the other hand, does not predict statistically significant facilitation, although numerically is in the correct direction ($M_{\text{Related}} = 0.73, SE = 0.02, M_{\text{Unrelated}} = 0.7, SE = 0.02, t(39) = 1.01, p = 0.31, d = 0.16$). This result is in constrast with the ones reported in Jones et al. (2006) and below (§3.6) we discuss some of the reasons of this difference. Random Indexing vectors fail to predict facilitation for the long-distance mediated pairs ($M_{\text{Related}} = 0.74, SE = 0.02, M_{\text{Unrelated}} = 0.73, SE = 0.02, t(39) = 0.75, p = 0.46, d = 0.12$). HAL representations also fail to predict facilitation ($M_{\text{Related}} = 0.67, SE = 0.03, M_{\text{Unrelated}} = 0.65, SE = 0.03, t(29) = 0.82, p = 0.41, d = 0.15$), whereas LSA vectors manage to reach marginal significance ($M_{\text{Related}} = 0.18, SE = 0.03, M_{\text{Unrelated}} = 0.12, SE = 0.02, t(39) = 1.8, p = 0.08, d = 0.29$.

McKoon & Ratcliff (1992)

McKoon & Ratcliff (1992, Experiment 3) argue for compound cue theory in semantic priming (Ratcliff & McKoon, 1988), an account which is directly related to the DSM mechanisms presented in Chapter 2. In short, within compound-cue theories when processing a word, semantic memory is accessed using a cue consisting of both the target word and the context in which it occurs (e.g., either the preceding word or the words in a window around the target). Since semantically related words tend to co-occur more frequently than do unrelated words, the compound cues for related words exhibit greater familiarity than do those for unrelated words. The above models all capture this either by fusing the target with its context via circular convolution (BEAGLE), by explicitly (HAL, LSA) or implicitly tracking the probabilities (neural embeddings).

In their study, McKoon & Ratcliff (1992) traverse the 1988 version of the Associated Press newswire corpus, and for each word in the corpus, they define a six-word window around it as context and measure the mutual information between the target and the context words. Mutual information is a measure of dependence between two variables telling us how much do we know for one variable given another. Concretely, it measures the probability of seeing two words in the same window divided by the probabilities of those words (3.3). This way of defining co-occurrence is very close to the way paradigmatic models describe co-occurrence as they yield the same output as HAL with a slightly altered orthographic frequency normalisation constant.[12] If the ratio in (3.3) is greater than one, then the words co-occur in the same window more than expected (i.e., more than it would have been expected if we simply encountered the words in the corpus). We can then measure whether this co-occurrence is significant using a t statistic (Ward Church & Hanks, 1989). Using this methodology, McKoon & Ratcliff (1992) create four lists of prime words; *semantically related*, words that have a strong chance of co-occurring in the same window (as captured by high t-values), words with a milder chance, albeit still significant, of co-occurring (low t-values) and unrelated. Using a lexical decision task, they find that both the *semantic* and *high t-value* conditions were significantly faster than the baseline, while *low t-value* was faster without reaching significance ($p = 0.09$),

$$\log_2 \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \tag{3.3}$$

where $P(w_i)$ is the probability of word i in the corpus and $P(w_i, w_j)$ is the joint probability of seeing both word i and word j in the window.

[12]The orthographic frequency normalisation method in Table 2.3 should read here $M_{ij} = \frac{M_{ij}}{\text{freq}(M_i)\cdot\text{freq}(M_j)}$

Table 3.9 Data from McKoon & Ratcliff (1992) and corresponding model predictions. The human data show RTs in each condition, whereas for the computational models we report the average cosine similarity between the *prime* and the *target* (the numbers in parentheses denote standard error of the mean).

Model	Semantic	High t-value	Low t-value	Unrelated
Humans	500	528	532	549
BEAGLE	0.87 (0.02)***	0.83 (0.02)*	0.82 (0.02)	0.78 (0.02)
COALS	0.72 (0.02)***	0.55 (0.03)***	0.5 (0.02)***	0.37 (0.03)
HAL	0.79 (0.03)**	0.73 (0.03)*	0.72 (0.03)	0.67 (0.03)
LSA	0.41 (0.04)***	0.23 (0.03)***	0.19 (0.03)*	0.1 (0.03)
Neural Embeddings	0.5 (0.02)***	0.32 (0.02)***	0.26 (0.02)*	0.16 (0.01)
Random Indexing	0.93 (0.01)***	0.91 (0.01)†	0.9 (0.01)	0.88 (0.01)

Significance levels: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$
Note: Significance testing was performed comparing the three groups against the unrelated condition (using paired t-tests). See text for more comparisons.

The neural embeddings predict facilitation in all groups (Semantic: $t(39) = 16.37, p \approx 0, d = 2.62$, high t-value: $t(39) = 8.59, p \approx 0, d = 1.38$, low t-value: $t(39) = 2.61, p = 0.01, d = 0.42$) compared to the unrelated controls. Furthermore, an one-way Analysis of Variance (ANOVA) with `group` as a between-subjects factor shows that the three groups varied significantly $F(2, 117) = 16.88, p \approx 0, \eta^2 = 0.22$. Subsequent pairwise comparisons using Holm's correction show that all three groups differed either significantly or marginally significantly from one another (Semantic > high t: $p \approx 0$, semantic > low t: $p \approx 0$, high t > low t: $p = 0.06$).

BEAGLE, similar to the behavioural results, predicts facilitation only for the first two conditions (Semantic: $t(39) = 3.74, p \approx 0, d = 0.6$, High t-value: $t(39) = 2.31, p = 0.02, d = 0.37$) but not for the low t-value condition ($t(39) = 0.72, p = 0.47, d = 0.12$). A between conditions one-way ANOVA was significant ($F(2, 117) = 3.53, p = 0.03, \eta^2 = 0.06$) revealing differences only between the Semantic and Low t-value conditions ($p = 0.03$).

COALS predicts faciliation in all three groups (Semantic: $t(37) = 11.21, p = 0, d = 1.84$, High t-value: $t(37) = 6.02, p = 0, d = 0.99$, Low t-value: $t(32) = 4.25, p < 0.001, d = 0.75$). An one-way ANOVA showed significant differences between the three groups ($F(2, 106) = 10.52, p = 0, \eta^2 = 0.17$) with the Semantic group differing from both the High and Low t-value conditions ($p = 0.01$ and $p \approx 0$, respectively).

HAL vectors predicted a pattern similar to the behavioural results, where only the first two groups were significantly different from the unrelated baseline (Semantic: $t(32) = 3.2, p < 0.01, d = 0.57$, High t-value: $t(34) = 2.56, p = 0.01, d = 0.44$, Low t-value: $t(33) = 0.86, p > 0.05, d = 0.15$). However, a subsequent one-way ANOVA did not reveal any differences in the means of each group ($F(2, 99) = 2.52, p = 0.09, \eta^2 = 0.05$), and none of the pairwise comparisons came up significant.

Similar to COALS, LSA representations predict facilitation in all groups (Semantic: $t(39) = 8.28, p = 0, d = 1.33$, High t-value: $t(39) = 4.02, p \approx 0, d = 0.64$, Low t-value: $t(39) = 2.43, p = 0.02, d = 0.39$). Furthermore, there were significant differences $F(2, 117) = 8.5, p = 0, \eta^2 = 0.13$ between the Semantic and both High and Low t-value groups ($p = 0.01$ and $p \approx 0$, respectively), but there was no difference between the two t-value groups ($p = 0.34$).

Random Indexing, on the other hand, predicts significant or facilitation only for the first group (Semantic: $t(39) = 4.09, p \approx 0, d = 0.65$, High t-value: $t(39) = 1.58, p = 0.12, d = 0.25$), but not for the low t-value condition ($t(39) = 0.44, p = 0.66, d = 0.07$). There were minor significant differences between the three groups as revealed by an one-way ANOVA ($F(2, 117) = 3.43, p = 0.04, \eta^2 = 0.06$), but only the difference between the Semantic and the Low t-value group was significant ($p = 0.03$).

3.6 Discussion for Studies 1 and 2

The results obtained from the above studies show that distributional models of semantics can account for unique variance in the reaction times of semantic priming tasks. The predicted facilitative effect of semantic distance between the prime and the target persists even when we control for different psycholinguistic variables such as word frequency and word length, and take care of issues of overfitting the model parameters. We see three ways the above results provide a contribution to the current literature. Firstly, they replicate effects observed in past studies obtaining better results through a more sophisticated model selection procedure. Secondly, they extend previous results using a more robust baseline model (i.e., WordNet). Thirdly, they indicate that we can recover unconscious 'semantic' representations with high precision. While, in the context of the present thesis, our primary concern is in the latter, we explore this aspect in depth in §4.7 having obtained results from cross-linguistic comparisons.

We achieve similar model fit to Mandera et al. (2017) using a more sophisticated parameter selection procedure and a much smaller corpus. Regarding the first point, we have already noted that DSMs come with many free parameters that can, potentially, alter the predictions of the models in a particular task. The number of possible partitions of the hyperparameter space grows exponentially with the number of possible values for each parameter rendering

parameter fitting in the present context particularly hard. Much of the problem is alleviated either by constraining the parameters we look at or by adopting a selection procedure that does not traverse the entire space. While Mandera et al. (2017) opt for the former examining solely the size of the embedding space and the context size, we choose the latter using the Bayesian Optimisation procedure outlined in §3.3.1. This choice permits us to converge faster to the best fitting model without constraining the possible values for the parameters. Secondly, the size of the corpus used in the above simulations is a fraction of the corpus used by Mandera et al. (2017). While the size of both corpora probably extends the number of words encountered in a lifetime (McDonald & Shillcock, 2001), we should keep in mind that humans can enrich their semantic representations using different sources. Furthermore, since reducing the size only improves/does not affect the results we conjecture that humans do not need massive amounts of information to form such representations.

3.6.1 Reliability Issues

While from a computational point of view, the above results might seem somewhat impressive in capturing the underlying organisation of the semantic memory, one needs to be cautious in concluding that such models can reliably predict semantic priming effects on the item level. Heyman, Bruninx, Hutchison & Storms (2018), for example, performing both original experiments and re-analyses of previously reported results, conclude that when examining the results on the *item* level (as opposed to *participant*) the reliability of the priming effects was quite low. While we consider this a valid point, the studies presented in this chapter did not aim at providing the optimal way of modelling semantic priming tasks, which can be marred by a myriad of confounds (Hutchison, 2003b), but to indicate that across several different experiments, which share a similar underlying structure, the best computational abstraction of how words are organised in the semantic memory.

The performance of WordNet

Despite its popularity within the NLP community (Fellbaum, 1998), the low performance of WordNet in the above tasks is notable as it was originally conceived as a model of semantic memory (Miller & Fellbaum, 1992). Given that and its ability to predict human semantic similarity judgements (Budanitsky & Hirst, 2006), it would be reasonable to expect for WordNet to be able to predict semantic priming effects, too. We argue that the poor performance of WordNet can be accounted for on two grounds; (a) the level of description it aims to explain and (b) its constrained view on the different word senses. We note in §1.3.1 that WordNet captures how concepts are represented in the mind irrespective of their linguistic realisation.

However, both the automaticity of the response achieved by the short SOA as well as the effect of surface psycholinguistic variables such as *orthographic frequency* or *word length* (see, §A.4) indicate that the linguistic realisation is of importance in this context as participants rely more on such information than on 'deeper' representations. Conversely, the ability of WordNet to model off-line similarity judgements can be explained in that participants can go beyond the given words and reflect on how the underlying concepts are linked.

We see two potential counterarguments to the above proposal, both methodological in nature. Firstly, we can attribute the relatively poor performance of WordNet to the synset selection procedure. Despite our best efforts to match the WordNet synsets to the words used in the SPP, our automated selection process was far from perfect. As described in §3.3.2, for each word pair in the related condition, we select the two synsets where similarity is maximised, whereas for the unrelated condition we randomly select two synsets. It would not be unreasonable, therefore, to conjecture that by not selecting the intended senses we either over- or under-estimate the priming effects. While this is a reasonable objection, it is hard to consider an alternative method without manual selection of the synsets. However, we note here that looking solely at the related pairs, where the method was motivated by the *spreading activation theory* the correlation between similarity estimates and priming effects was not significant ($r(427) = -0.05$, $p = 0.25$).

The second explanation of the poor performance of WordNet in the above tasks would be that we did not perform any hyperparameter tuning. The lack of free parameters renders WordNet a very simple, from a statistical point of view, model. DSMs, on the other hand, increase their degrees of freedom by having several hyperparameters. Tuning these hyperparameters can result, albeit not necessarily, in a better fit to the data. The statistical analyses performed above show that DSMs provide a better alternative to WordNet on this sort of data as the increased complexity is justified by the statistically higher R^2 values. Despite the better fit, being able to fit the DSMs to this particular task, whereas we are bound to use vanilla WordNet representations is –at least on some level– biasing against WordNet, the same way that Hutchison et al. (2008) biases against LSA.

While we acknowledge that such alternative explanations exist, we still maintain that the lower fit of WordNet can be attributed to the fact that it captures a qualitatively different aspect of semantics from DSMs. To further support this notion consider two similarity judgement tasks in which participants rate the similarity between two concepts (Miller & Charles, 1991; Rubenstein & Goodenough, 1965). Budanitsky & Hirst (2006) found that the best fitting WordNet similarity metric (in that case the Leacock-Chodorow metric) achieves a correlation of 0.84 and 0.81, respectively, to the human subjects. The best DSM achieves slightly lower correlations of 0.79 and 0.8, respectively. To explore whether the differences between the two

similarity estimates are statistically significant, we construct two linear models where the human similarity score is the dependent variable and the similarity estimate the independent. In all the assessment procedures we outline in §3.4.3 the WordNet similarity metrics provide a better fit to the human data ($AIC_{WordNet}$ = 222.88, AIC_{NN} = 240.25, B_{wn} = 4157).[13] The results from the offline similarity judgement studies, together with the statistically significant correlation between the WordNet estimates and the priming effects from the SPP in the long SOA condition suggest that WordNet might model a conscious aspect of semantic knowledge.

The issue of multiple senses

Unlike WordNet, DSMs conflate all the possible senses words can have. Take the word *bank*, for example, which can either mean (1) the land alongside a river ($bank_1$) or (2) the financial establishment ($bank_2$); from the point of view of the DSM, there are no different senses, just the same four byte string. Presumably, if the words were somehow marked to denote different semantic senses, then $bank_1$ would have different distributional patterns than $bank_2$. Recovering the intended sense of a word from its distributional patterns is a quite active area of research (Iacobacci, Pilehvar & Navigli, 2016; Pilehvar, Jurgens & Navigli, 2013; Rothe & Schütze, 2015, for some recent developments), however, we choose to employ the simpler approach which does not distinguish between senses. We motivate this decision by arguing that sense information is not available in the 200ms SOA condition. Till, Mross & Kintsch (1988), for example, performed a study in which participants read short text passages containing a *prime* and immediately after reading the passage they had to perform a lexical decision on a target word that appeared on the screen. This target word could either be an associate of the prime or a probable inference suggested by the text. It was found that at shorter SOAs (<400ms), only the associates showed facilitation compared to unrelated control words, whereas at longer SOAs inference words were also facilitated. These results suggest that sense selection is *post*-lexical, hence not taking place during the semantic priming task (at least in these conditions). Such results have also been incorporated in models of discourse representation (Kintsch & Mangalath, 2011) that use a single DSM representation for each word in the corpus which then takes a specific sense by the words that appear in the same propositional structure. Furthermore, a similar point could be made from priming from homophones (e.g., Seidenberg, Tanenhaus, Leiman & Bienkowski, 1982) where both senses are primed when the target occurs immediately at the offset of the word, but there is only priming of the contextually relevant sense at an interval of 200ms (from the prime offset). However, one should be cautious in

[13] B_{wn} is the probability of the data under WordNet relative to the neural embeddings.

relating the results from auditory priming studies to those of visual priming, as it is harder to equate the 200ms interval in those studies to the SOA in the visual priming studies.

BEAGLE

More notably, BEAGLE failed to predict significant facilitation for the mediated priming effects, although numerically it was in the correct direction. Jones et al. (2006) report that BEAGLE can accurately predict the effects in the three studies presented above. As we mentioned in §2.2.3, BEAGLE combines two independent streams of contextual information. *Ordering*, which similar to HAL looks at a window around the target word, and *contextual*, which similar to LSA forms representations by using the entire document (here, sentence) as context. Having computed those, it performs a linear translation (2.21) shifting the coordinates of the target word in one space by the direction it has in the other, forming thus, combined memory representations. Jones et al. (2006) report the simulations in all three representations (i.e., order, context and both) and remark that different streams can drive the predictions on various phenomena. For example, in Chiarello et al. (1990) the *order* vector drove the semantic condition predictions. For the mediated priming effects, Jones et al. (2006) find that the *context* vector drives the results reporting small effects, if any, for the *order* representations.

However, since the *context* vector is the one chosen from the optimisation procedure, we would expect the predictions in the mediating priming to be closer to the behavioural results. On the other hand, the chosen optimal dimensionality was 256 units, well below the 2048 units used in Jones et al. (2006). We, therefore, argue that the poor predictions on mediated priming for BEAGLE can then be accounted for when we consider what this model was trained to maximise. Presumably, the optimal solution for the naming task did not require more information than a compressed version of the document context, something that was not sufficient for mediated relations to be made apparent in the semantic vectors.

LSA

A question that arises from these results concerns the failure of LSA to enter as a significant factor in the analysis reported in Hutchison et al. (2008). The difference between our formulation and that study is that we change the corpus to a considerably larger one, using vectors of the same dimensionality. Although as a factor LSA was weak and its by-relation correlations rather poor, it still enters as a significant predictor in our regression model. Hutchison et al. (2008) do not provide any information regarding their training regime (e.g., the corpus LSA was trained on or any further details). Based on past studies, we conjecture that they used the vanilla LSA vectors originally made available by Landauer & Dumais (1997) which were trained

on the TASA corpus (≈ 11 million words). An explanation, therefore, for the relatively poor performance in that study would be that the BNC is a qualitatively better corpus providing richer (and more relevant) LSA representations than TASA.

Mediated Priming

Given the discussions in §§ 2.2.2, 2.3.1 and 3.2.1, we identify two problems related to the results of the mediated priming studies in §3.5. Firstly, how is HAL able to predict facilitation for the *Semantic* and High *t*-value conditions in McKoon & Ratcliff (1992) contrary to the results obtained by Livesay & Burgess (1998) presented in §3.2.1? Secondly, how is the neural network able to capture mediated priming effects? In our presentation of neural embeddings §2.3, we saw that their goal is to find a lower dimensional description of the target word's contextual distribution where context is defined as *n* words before or after the target. From this point of view, the neural embeddings are closely related to paradigmatic models (e.g., HAL) which –purportedly– cannot capture mediated priming effects.

Let us look more closely at the first question of why HAL has moderate success in predicting the effects in McKoon & Ratcliff (1992). We will look only at the two co-occurrence conditions (high *t*-value and low *t*-value) and not the semantic one, as HAL is expected to be able to predict the effects there (e.g., when the prime-target pair is *finger–hand*). As we remarked above, McKoon & Ratcliff (1992) traverse a corpus of American English and record for each word the mutual information between that word and every other word that appears in its context, defining context as *six* words around the target (i.e., three before and after). As seen in (3.3), the mutual information metric is symmetrical between the two words. That is, if *tiger* appears in the context of *stripes* ten times, then *stripes* is going to appear an equal number of times in the context of *tiger*. When HAL is, therefore, building the vectors for those two words, there is going to be an overlap between the words that commonly co-occur with *tiger* and *stripes*. In the High *t*-value condition, the overlap should be greater as the two words co-occur more with each other, whereas this effect should be lessened in the Low *t*-value condition.

While the above explains the behaviour of HAL in McKoon & Ratcliff (1992), it fails to account for the patterns observed in Balota & Lorch (1986). We argue that this dissociation stems from the fact that the way McKoon & Ratcliff (1992) conceptualise mediation is qualitatively different from Balota & Lorch (1986) as it is too restrictive. Take an example of a mediated pair from Balota & Lorch (1986); *war* and *quiet*, where *war* is related to *peace* and *peace* to *quiet*. Defining a window of six words around *war*, its mutual information to *quiet* is 0 as there are no contexts for *war* where *quiet* appears. Indeed, we recorded the mutual information for every word pair in the two studies using the method by McKoon & Ratcliff (1992) and display the results in Table 3.10. As a comparison, we also calculated the mutual

Table 3.10 Average Mutual Information using two different methods for calculating it between the word pairs in McKoon & Ratcliff (1992) and Balota & Lorch (1986). The window-based approach defines context as six words around the target (three before and after), whereas the document-based approach considers as context the entire sentence in which the two words appear.

Study		Window-based	Document-based
Balota & Lorch (1986)		0.82	5.29
McKoon & Ratcliff (1992)	High *t*-value	3.58	6.28
	Low *t*-value	1.77	5.83

information using an approach more similar to the syntagmatic models. This is, instead of defining as context several words before and after the target, we define context as the entire document (here, sentence) in which the word appears. The results obtained shed some light on how LSA can predict facilitation in mediated priming, while HAL is able only to do so up to a certain extent. If we take mutual information as an index of word similarity (making the additional assumption that words that appear in each other's contexts will be more similar), then a document-based approach will place the mediated word pairs closer together, whereas the window-based only partially (for the High *t*-value condition). Paradigmatic models such as HAL are then expected to be able to predict facilitation insofar as the mediated pairs appear in the contexts enough times, something that is neither a necessary or a sufficient condition of mediated priming.

We now turn to the second question about why the neural embeddings, while using a window-based approach can predict the facilitation in the mediated condition in Balota & Lorch (1986). We argue that this behaviour is related to the way the neural network performs dimensionality reduction when predicting the contextual distribution for each word. Recall that to form the word representations the network predicts the words that can appear in the target's window. The errors then are used to update the lower dimensional description of this distribution (i.e., the word representations). Take the words *tiger* and *lion*, for example; the network will try to predict the word *stripes* in the contexts of *tiger* but not to those of *lion*. However, when processing the word *tiger* and the errors are propagated back to the lower dimensionality layer, no single unit encodes the word *stripes*. What the network will do, therefore, is to distribute the errors in such a way such that the embeddings capture the most variance of the output layer (i.e., the words that *generally* appear in the contexts of *tiger*). Since the embeddings for *tiger* and *lion* are quite similar, they will activate similar nodes in the output layer. Mediated priming is, therefore, accounted for as the embedding for *lion*

residually activates the node for *stripes* in the output layer only because it shares a contextual distribution with *tiger*. This inferential process explanation is in line with the discussions offered by Landauer & Dumais (1997) and Shepard (1980) where they view dimensionality reduction as a way of performing inductive inference.

The results of the present study are quite positive in that DSMs provide an accurate approximation of the semantic memory. While some of the models have specific issues related to how they define context, they are all quite successful not only in predicting the priming effects directly but also in modelling other studies of semantic priming which explore more fine-grained details of the semantic memory. The best performing model in all simulations seems to be the neural network as not only it yields the highest model fit and does so by better modelling the different semantic relations but also because is able to account *mediated* priming phenomena in a human-like way. We now proceed to examine whether these results hold in languages other than English. In §4.1 we outline some of the reasons this class of models might fail in predicting the effects in other languages something that is of utmost importance when we model semantic implicit learning tasks in Chapter 5.

Chapter 4

Cross-linguistic exploration of the unconscious representations

4.1 Introduction

The studies reported above focus on the ability of VSMs to capture pairwise relations between lexical items. While they can do so in a human-like way, they might be under-informative as to whether they managed to capture aspects of the general structure of the semantic memory (Buchanan et al., 2001; Mirman & Magnuson, 2008; Schreuder & Flores D' Arcais, 1992). Are there, for example, regions in the semantic space where there is more concentration of concepts? Can this affect processing? Questions such as these are difficult to answer only by looking at pairwise relationships such as the above and require a more general approach to the problem. Recent studies (Buchanan et al., 2001; Chen & Mirman, 2012; Mirman & Magnuson, 2008; Shaoul & Westbury, 2010; Siakaluk, Buchanan & Westbury, 2003) have widened this scope by exploring the effects of *semantic neighbourhood density* (as captured by those models) in the early stages of word recognition as an analogue to orthographic and phonological density. The advantage of examining the effects of semantic neighbourhood density in early word recognition is that it takes into account the position of the word in this high-dimensional space relative to a much larger sample of words instead of a single one.

The effect of neighbourhood density is pervasive in lexical decision tasks often yielding disparate results. In one case, Sears, Hino & Lupker (1995) in a series of experiments found that the number of orthographic neighbours can have a facilitative effect on *written* word recognition and Yates (2005) found a similar effect for phonological neighbours. However, these effects seem also to be modulated by word frequency (Andrews, 1992; Grainger, O'Regan, Jacobs & Segui, 1989). On the other hand, Luce & Pisoni (1998) found that in *spoken* word

recognition the same variables (phonological and orthographic neighbourhood density) can have an inhibitory effect. This interaction is explained at the perception level; spoken words are perceived serially (Marslen-Wilson, 1987). Hence, more phonological neighbours at the onset would increase entropy and slow down recognition. Conversely, written word recognition is parallel (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982), hence, whatever disambiguating information orthographic or phonological neighbours can provide is available from the beginning (Chen & Mirman, 2012).

Only because measures of orthographic and phonological density are easier to derive one cannot assume that they are the only ones that can have an effect on word recognition. Consider an attractor network such as those described in §3.2.2. In principle, in such a model, words with semantically denser neighbourhoods (i.e., words that share semantic features with more words) should be easier to recognise than words with fewer neighbours. The reason for this facilitative effect is because more neighbours can *pull* the gradients towards the correct attractor (in §4.6 we discuss the seemingly diverging results of Mirman & Magnuson, 2008). Simply put, there are more *entry* points for the right pattern than for a word with fewer neighbours.

For this study, we look at Lexical decision reaction time (LDRT) data. There are two reasons behind this choice; firstly, semantic density plays a more important role in lexical decisions than in naming tasks (Buchanan et al., 2001; Neely, 1991). We can attribute this effect to the high spelling-sound consistency in naming tasks (Plaut, McClelland, Seidenberg & Patterson, 1996) meaning that semantics has a weaker effect (explaining the lower scores obtained above). Secondly, another advantage in taking LDRTs as our dependent variable is that they permit us to look at languages beyond English. In the last few years, several large-scale studies have appeared recording LDRTs for French, Dutch, Malay and Chinese (explained in more detail in §4.3.1) along with some item-level variables for the words used (e.g., frequency or orthographic neighbourhood density). Apart from further corroborating the above results, we can also explore whether DSM predictions are only sensitive to properties of English or are more 'universal' (Qian, Qiu & Huang, 2016; Upadhyay, Faruqui, Dyer & Roth, 2016). Since semantic implicit learning has been reported in languages other than English (Chen et al., 2011; Leung & Williams, 2014; Pastorino Campos, 2017), these results form the basis for the studies reported in §§ 5.4 and 5.6 where we compare the predictions of semantic implicit learning models against data coming from speakers of English and Chinese.

Although DSMs have been applied successfully in languages other than English (Qian et al., 2016; Upadhyay et al., 2016), we see two potential problems regarding their ability to capture unconscious representations. Firstly, languages such as French and Dutch have richer inflectional morphology than English. From an estimation point of view, this results in higher

type-token ratios (i.e., more words associated with a single lemma), which, in turn, has two effects; (a) each token in the corpus being encountered fewer times, and (b) both sparser and longer contextual probability distributions. Regarding (a), the problem is that each word is not 'seen' in as many contexts to reliably predict them, and, as for (b), there are more potential contexts in which each word can appear.

We could resolve this issue by stemming the corpora retaining only the lemmas instead of the word tokens. That is, we replace all occurrences of *going*, *went*, and so on, with the string *go*. Retaining only one form instead of many both increases the times we encounter each word and reduces the size of the probability distribution to be estimated. Regarding LDRTs, however, Moscoso del Prado Martín, Kostić & Baayen (2004) note that 'aggressively' discarding word tokens leads to *over*-estimating each word's frequency providing a worse predictor of LDRTs. To exemplify this *over*-estimation, consider the example of *go* (also given in, Bentz, Alikaniotis, Samardžić & Buttery, 2017b) where its inflectional variants are; *go* 10, *going* 6, *went* 3, *gone* 2, *goes* 1, and *goeth* 1. If we consider all these forms variants of the word *go* and collapse their counts, *go* would receive a frequency of 23 instead of its original 10.

Secondly, in §2.2 we made a distinction between *paradigmatic* and *syntagmatic* vector-space models. We argued that different semantic relations are captured by different types of models. For example, relations which we consider as semantic (that is, words which belong to the same class) are *paradigmatic* (Firth, 1957). The issue with the above is that when we look at different languages, it is unclear which relations should be considered paradigmatic and which syntagmatic. Cruse (1986), for example, argues that antonymic relations are paradigmatic in English as in Exs. (4.1) and (4.2),

(4.1) he is big.

(4.2) he is small.

(4.3) * he is big small.

On the other hand, Ex. (4.3) is perfectly legal in Chinese where antonyms can appear next to each other forming generic terms (Li & Thompson, 1981). For example, the collocation 大小 ('big-small') in Chinese refers to the generic term *size*. While we cannot claim that participants' RTs are affected differently because of a different kind of relation, since different models target one of the two types of relations, we would expect differing predictions across languages.

4.2 Semantic Neighbourhood Density

4.2.1 Prior work

The study of the effect of semantic neighbourhood density on lexical access stems from the work of Buchanan, Hildebrandt & MacKinnon (1994) on patients with deep dyslexia. Buchanan, Hildebrandt & MacKinnon (1999) compared the predictions of two models of word reading on three patients with deep dyslexia. The authors reasoned that in deep dyslexia, semantic neighbourhood *size* would have an inhibitory effect on lexical access. According to the authors, in deep dyslexia patients name the incorrect words because they are unable to suppress spurious activations of phonologically related words. Consequently, semantic neighbours would only amplify this effect yielding even worse performance. The authors used semantic neighbourhood density measures derived from word association experiments on 218 undergraduate students but failed to find a link between their estimates and performance in naming tasks. Based on the results from Lund et al. (1995) we described above, they reason that these deficits are more semantic than associative. Hence, this effect would not manifest in word association norms but should be captured by an HAL-like model (which favours paradigmatic relations). Indeed in subsequent simulations reported in Buchanan, Burgess & Lund (1996) and Buchanan, Kiss & Burgess (2000), Semantic Neighbourhood (SN) measures derived from HAL were able to predict naming times for impaired readers.

Buchanan et al. (2001) explore the effect of semantic neighbourhoods on *word recognition* using both lexical decision and naming tasks. Using HAL vectors trained on a USENET corpus, they derive SD values for each word, which they enter in a hierarchical regression model along with *orthographic frequency*, *orthographic neighbourhood* and *word length*. Buchanan et al. (2001) find that *semantic density* was a significant predictor of LDRTs even when more conventional measures, such as imageability ratings, were entered in the model. The slope of the semi-partial correlation between RT and SD indicates a facilitatory effect of the semantic neighbourhood size in the lexical decision (i.e. the larger the neighbourhood, the lower the RT). While these results might seem at odds with the above, they concern LDRTs gathered from participants who did not report any cognitive impairments. By analogy to the rest of neighbourhood measures described above, semantic neighbourhood measures also show a facilitatory effect in word reading. Subsequent experiments give further support to these initial results (Balota, Cortese, Sergent-Marshall, Spieler & Yap, 2004; Shaoul & Westbury, 2006, 2010; Siakaluk et al., 2003; Yates, Locker & Simpson, 2003).

Mirman & Magnuson (2008) attempt a more fine-grained analysis of the effects reported in the above studies. To explain the diverging results obtained by using different methods of calculating semantic neighbourhood density, they reason that there might be a difference

for words that have many *near* neighbours as opposed to words that have many *distant* neighbours. Using the McRae norms, they compile stimuli lists using a 2×2 factorial design (size of distance neighbour set vs. size of near neighbour set) and find that neighbourhood size is inhibitory for the words with many *near* neighbours and vice-versa for the words with many *distant* neighbours. They operationalise the distinction between near and distant neighbours by taking cutoff points in the cosine distance between the target word and all the other words (e.g., words that have a cosine distance of .5 or greater to the vector of cat are near neighbours). The dual role of Semantic Neighbourhood Density (SND) poses problems to traditional models of word recognition which assume that neighbourhood density is going to have either a facilitatory or inhibitory effect. Instead, Mirman & Magnuson (2008) (also in Chen & Mirman, 2012) explain their results in terms of a model based on attractor dynamics (see §3.2.2). More specifically, while *all* neighbours create a gradient that facilitates settling to the correct attractor, for *near* semantic neighbours $\cos(\hat{\mathbf{y}}, \mathbf{y})$ is going to be high for many $\mathbf{y}$'s creating competition among the candidates.

The aim of the two studies presented in this Chapter is threefold. Our main objective is to ensure that the DSMs used in the previous Chapter are sensitive to the semantic neighbourhood density effects found previously in the literature. If we are successful in finding SND effects using DSMs we can validate the results from the previous studies using a measure that takes into account the global configuration of the semantic space instead of pairwise relations. Furthermore, we will be able to extend previous results on a large dataset of LDRTs instead of small-scale studies. Finally, it will enable us to explore whether we can obtain similar effects in languages other than English where we have large datasets of LDRTs but not of semantic priming effects.

4.3 Method

In geometric approaches to meaning, the *size* of the semantic neighbourhood of a target word w_t is the number of words that exist within a fixed distance from w_t. Essentially, we construct an n-dimensional manifold to enclose all the words with which w_t is associated. Consequently, semantic neighbourhood *size* is the number of word-points within the sphere and *density* the average distance between the w_t and all the other words in its neighbourhood. Concretely, we compute the distance between a word w_t and all the other words in the vocabulary using any of the metrics in Table 2.4. Then, we sort the values of this vector in decreasing order. The problem is that there is no principled way of deciding the radius of this manifold, that is, the threshold which divides neighbours from non-neighbours.

A simple approach proposed by Buchanan et al. (2001) is to average the distances between the corresponding word and its N closest neighbours (usually 10) and then use this *density* measure to *infer* the size of the neighbourhood. A dense neighbourhood, one where the closest words are tightly packed around the target word, implies large semantic neighbourhood, whereas, thinly scattered close words imply a low neighbourhood size. Again, the cutoff point chosen is arbitrary, and although it could be left as a free parameter a value of 10 has been found to yield good results in a number of studies (Balota et al., 2004; Siakaluk et al., 2003; Yates et al., 2003).

One issue with the above approach noted by Shaoul & Westbury (2006) is that it assumes neighbours are evenly distributed around the target word. This relatively strong assumption can potentially mask the *within-neighbourhood* variability in cases where two words have entirely different neighbourhood distributions. Shaoul & Westbury (2006) propose to fix this issue by asking whether two words stand closer or further away in this semantic space than any two words would be in the corpus. The idea is that if the similarity between a particular word pair is higher than what it would have been expected given a matrix of similarities, then two these words have to be neighbours.

To formulate this problem, Shaoul & Westbury (2006) traverse the word similarity matrix gathering a lot of pairwise similarities between word-vectors. By finding the mean and the standard deviation of this sample, they consider as neighbours words that were more similar to each other a number of standard deviations above the average similarity any two words have in the matrix. Concretely, let $\mathbf{s} \in \mathbb{R}^K \sim \text{vec}(\mathbf{S})$ be a vector of K samples from the vectorised triangular similarity matrix $\mathbf{S} = \mathbf{M}^\top\mathbf{M}$. (a large K can approximate the global similarity estimate).[1] Now let $\mathbf{s}' = \frac{\mathbf{s}-\bar{\mathbf{s}}}{\sigma}$ be the vector of standardised z-scores for $\mathbf{s}$. We can use these standard scores to find whether a similarity value is higher than expected from the word matrix. Setting an arbitrary cutoff point for standard deviations above the mean t (e.g., 1.5SDs above the average) we can find the neighbourhood size and density as the size and mean of the following set: $\{s_i,\ \forall i \in \mathbf{s} | s'_i \geq t\}$.

For example, if the average similarity in the corpus is 0.3 and 2 SDs above the mean is 0.55, then for each word, its neighbours are the words, with which its similarity estimate is above 0.55. While in the above method there is no psychological grounding on how to find the optimal threshold value, heuristically, a value 1.5 – 2 SDs above the mean yields the best results regarding model fit Shaoul & Westbury (2010). Since this method essentially finds the radius of the n-sphere, within which a word can be considered as a neighbour, we derive two new metrics: the size of the neighbourhood –Number of words within a radius (NCOUNT)–

[1] We use this notation mainly for clarity. The above formulation introduces a major computational bottleneck as computing $\mathbf{M}^\top\mathbf{M}$ requires $|\mathcal{V}|^3$ operations. In reality, the algorithm samples words from the vocabulary without replacement and computes the distance values there.

as found by the number of words within that distance and density of the neighbourhood found by averaging the distances between the target and its neighbours –Average Radius of Co-Occurrence (ARC)– (see above). Both of these metrics have been found to correlate highly with RT in a range of semantic tasks (Shaoul & Westbury, 2010).

One potential issue with this approach is that it masks *between-word* variability. That is, assuming that in high-dimensional spaces the distance between any two vectors is going to be maximal (this is a common issue in high-dimensional datasets, see, e.g., Radovanović, Nanopoulos & Ivanović, 2010) few highly correlated words can skew the above estimates. Ignoring this variability and choosing an absolute threshold value for all the words will bias the number of neighbours in favour of words with *very* close semantic neighbours nullifying any effect for all the other words in the vocabulary. Indeed, from a quick exploration of a publicly available dataset of 57153 NCOUNT scores[2] we see the 75% of the words have no semantic neighbours at all whereas for 270 words SN values are more than 9000.

In our simulations, we tested the proposal of Shaoul & Westbury (2006) along with averaging the distance to the ten closest words. In unreported simulations done on smaller portions of the British Lexicon Project (BLP) (to be explained below) we find that these two methods do not differ significantly from each other. Because of this, we focus solely on the ten nearest neighbours approach as it more intuitive and easier to compute (see, Appendix B, for more details).

4.3.1 Datasets

Semantic Priming Project

For the reported experiments on English we use the lexical decision task from the SPP. Apart from the task specification, deciding between words and non-words, the entire setting is identical to the description in §3.4.1. Non-words in the SPP were generated by changing one or two letters from the targets to form pronounceable non-words. While semantic priming effects can still be shown in the non-word case (Deacon, Dynowska, Ritter & Grose-Fifer, 2004), we focus only on the cases where the target was a word.

Lexicon Projects

A series of datasets called Lexicon Projects are collections of loosely-related projects which provide lexical decision data for a variety of languages. They generally contain LDRTs along with descriptive characteristics of the words used in the studies. To our knowledge, the following datasets exist:

[2] Available at `http://www.psych.ualberta.ca/~westburylab/publications.html`

1. The **English Lexicon Project** (Balota, Yap, Hutchison, Cortese, Kessler, Loftis, Neely, Nelson, Simpson & Treiman, 2007) which contains speeded naming and lexical decision for 40481 words and nonwords.

2. The **Dutch Lexicon Project** (Keuleers, Diependaele & Brysbaert, 2010) contains LDRTs for 14000 Dutch mono- and disyllabic words and the same number of nonwords.

3. The **French Lexicon Project** (Ferrand, New, Brysbaert, Keuleers, Bonin, Méot, Augustinova & Pallier, 2010) contains LDRTs for 38840 French words and 38840 pseudowords.

4. The **Malay Lexicon Project** (Yap, Liow, Jalil & Faizal, 2010) contains LDRTs for 9592 Malay words. Regarding their frequency estimates, those were derived newspaper corpora (ca. 7 million words).

5. The **British Lexicon Project** (Keuleers, Lacey, Rastle & Brysbaert, 2012) which contains lexical decision data for 14365 mono- and disyllabic English words and the same amount of nonwords.

6. The **Chinese Lexicon Project** (Sze, Liow & Yap, 2013) contains LDRT data for 25000 Chinese two-**character** compound words together with character frequencies based on subtitle corpora (Cai & Brysbaert, 2010).

since we use the SPP for English LDRTs, we focus on the French, Dutch, Malay and Chinese datasets.

4.3.2 Corpora

Here we describe the acquisition process and pre-processing of the corpora used to train the DSMs on French, Dutch, Malay and Chinese.[3] Unless otherwise stated, the pre-processing steps are similar to the ones we followed for the English corpus described in §2.4.2.

French corpus

For the French language simulations we use a portion of the frWaC corpus[4] (1.6 billion words) (Baroni et al., 2009) constructed from the Web. The texts were gathered by crawling the `.fr` domain and using medium-frequency words from the Le Monde Diplomatique corpus and basic French vocabulary lists as seeds. We reduce the training time by trimming the corpus into its first 10^8 words (size comparable to that of the BNC).

[3] In choosing our corpora our procedure was 'established available' corpus > 'subtitle' corpus > 'wikipedia' corpus.

[4] `http://wacky.sslmit.unibo.it/doku.php?id=corpora`

Dutch corpus

For the Dutch language simulations we use the Dutch portion of the 2016 edition of the OpenSubtitles corpus (Lison & Tiedemann, 2016). This provides an already cleaned corpus to which we can apply our pre-processing pipeline. Since the size of the original corpus is larger than the others, we restrict the size of the corpus by randomly selecting subtitles from movies which have been released after 2000, providing thus a more contemporary use of Dutch.[5] The final size of the corpus is thus reduced from 4.94×10^8 to 9.34×10^7 words.

Malay corpus

The lack of a Malay corpus prompted us to use the Wikipedia as a source of Malay texts. We use a Wikipedia dump file containing only the articles and article titles[6] totalling ca. 2.2×10^7 words. Prior to our usual pre-processing steps, we use a customised wikipedia extractor to clean the wikitext.[7] Moreover, we removed all lines less than ten words to ensure that we were not also considering the article titles, but we did not apply step 3 of our pipeline as Malay is an aggluntative language and words longer than 20 characters are to be expected.[8]

Chinese corpus

A problem we face when working with distributional semantics and Chinese corpora is that words are not delimited by a special character such as whitespace. Studies tracking the eye movements of native speakers of Chinese (Bai, Yan, Liversedge, Zang & Rayner, 2008) show that there is no advantage of space-delimited text. However, models of distributional semantics are based on the existence of separable units to form their vectors. A case in point is illustrated by (4.4) and (4.5) that show that a different segmentation of the same sequence of characters can give change the meaning of the sentence. While the second example is quickly discarded as semantically implausible, from the point of view of a DSM, it certainly remains plausible. Choosing one alternative over another will give rise to different estimates which in turn will return erroneous vectors. While as a problem this is related to the multiword expression problem discussed in §2.4.1, here the entire language is affected rather than a small subset of it.

[5] The original corpus contains movies from 1912.
[6] The dump file was generated on 1/9/2016.
[7] `https://github.com/dimalik/wiki-cleaner`
[8] In practice, however, they make up a negligible portion of our corpus ($\approx 0.04\%$).

(4.4) 日文 章魚 怎麼 說?
Japanese octopus how say
'How to say octopus in Japanese?'

(4.5) 日 文章 魚 怎麼 說?
Japanese article fish how say

(Xue, 2003)

There are three potential workarounds to this problem each with its own drawbacks; (1) obtain a large, unsegmented, corpus of Chinese such as a Wikipedia dump (ca. 10^8 "words") and use a state-of-the-art segmenter (Tseng, Chang, Andrew, Jurafsky & Manning, 2005) to transform it in a DSM-friendly format. In reality, this returns a very high Type-Token Ratio (TTR) (0.90) which we attribute to the irregular nature of Chinese Wikipedia articles (mix of traditional and simplified texts, potential mishandling of non-ascii characters). This TTR means that most words are not sufficiently encountered to form semantic vectors. (2) A simpler method would be to treat each character as a word and when needed use an algebraic method to combine them (such as addition or convolution) (see §2.4.1). While this method could be supported by the fact that many Chinese characters are words in their own right, it will be insufficient as most chinese characters can not only be words but also morphemes. (3) A third more elegant way would be to obtain already segmented corpora. While this might sound ideal, the size of those corpora is a fraction of that of the other languages. This can be a problem for the particular class of models we are looking at, and we train for more passes over the same corpus in order to increase the data the model encounters.

The corpus we ended up with is a concatenation of the four corpora used in the Second International Chinese Word Segmentation Bakeoff. Table 4.1 shows some statistics of the corpora which were used in the competition. These four corpora were concatenated in a single space delimited file[9] which after a preprocessing step similar to the ones employed in the other corpora was used as input to the DSMs.

4.3.3 Baselines

For the Chinese dataset, we retain the frequency of the *first* and *second* character (`C&B--Subtitle--CD--C1` and `C&B-Subtitle-CD-C2`, respectively) as well as the frequency of the entire word as estimated from two different corpora (`C&B--Subtitle--CD--W` and `Google--`

[9]The corpora include separate training and gold test sets. Prior to any concatenation we collated the training and the gold test sets.

Table 4.1 Summary of the corpora described in §4.3.2 and used in the simulations reported in §4.5.

Language	Number of Words	Source
Chinese	10758340	Second International Chinese Word Segmentation Bakeoff
Academia Sinica	5572191	
City University of Hong Kong	1496566	
Microsoft Research	2475264	
Peking University	1214319	
Dutch	93476724	OpenSubtitles 2016
English	97987371	BNC
French	92393223	frWaCs
Malay	22664164	Wikipedia

`freq--W`). The first corpus was based on film subtitles (Cai & Brysbaert, 2010) and the second was the number of entries for each word indexed in the Hong Kong traditional Chinese database in Google. Interestingly, the *lasso* feature selection retained both the subtitle frequency and the frequency of the word in the Google corpus. While this might seem as overfitting, the two different corpora are thought to capture distinct aspects of word usage; subtitle frequencies are considered to be a proxy of spoken language, while the Google corpus of written language (Brysbaert & New, 2009). As such, they provide two qualitatively different variables. Transforming all the variables to log-scale, the baseline model achieves a fit comparable to that reported in the original paper ($R^2 = 0.31, F(4, 5099) = 582.3, p = 0$) compared to $R^2 = 0.364$. Such differences, however, can be attributed to the fact that we drop words where the accuracy was not 100%.

For Dutch, we retain the word frequency as estimated from the CELEX corpus (H., Piepenbrock & Guilikers, 1995) (`celex.frequency`), the frequency of the word in the SUBTLEX-NL (`subtlex.frequency.million`), the length of the stimulus in characters (`nchar`), and the average orthographic Levenshtein distance of the 20 most similar words (`OLD20`). Our baseline achieves a $R^2 = 0.3, F(5, 3215) = 273, p = 0$. Our estimates are somewhat lower than those reported in the original paper ($R^2 = 0.34$), but such inconsistencies can be attributed to our different stimulus selection procedure.

For English, we use the same variables as in §3.4, keeping the *length* and the log-*transformed subtitle frequency* for the prime word and the *length*, log-*transformed subtitle frequency* and *number of orthographic neighbours* for the target. The baseline model achieves a fit $R^2 = 0.3$

overall, $R^2 = 0.29$ and $R^2 = 0.27$ for the validation and testing sets, respectively. Further to that, we also explore the estimate on the British Lexicon Project (described above). We do so as to ensure that our results are comparable to the other languages. In that dataset, we retain the *number of orthographic neighbours*, the *length* of each word, the *number of lemmas* that can be constructed from the target, the *frequency* of the word in the BNC and the contextual diversity as computed in the SUBTLEX–US corpus (Brysbaert & New, 2009). This fit of this baseline model was $R^2 = 0.193$.

For French, we retain the log-transformed frequency of each word as estimated by a subtitles corpus (`lcfreqmovies`), the number of letters (`nletters`) and the number of syllables (`nsyllables`). The baseline model achieves a high fit to the behavioural data ($R^2 = 0.38, F(3, 14496) = 3042, p = 0$). As a comparison, the non-standardised dataset achieves a substantially lower fit to the data ($R^2 = 0.28, F(3, 14496) = 1842, p = 0$). These values are close to the ones reported in (Ferrand et al., 2010) for the normalised and non-normalised data (35.1 and 32.4, respectively), albeit with a slightly different normalisation procedure.

For Malay we retain the number of letters, the number of orthographic neighbours, as well as the log-transformed frequency of each word based on a Singaporean corpus (`lg_freq_singapore`). Our baseline model achieves a fit comparable to that reported in the original paper ($R^2 = 0.55, F(3, 1508) = 609.9, p = 0$ vs. $R^2 = 0.555$). Note that in this model we did not log-transform the other variables (number of letters and orthographic neighbours).

4.4 Study 3: Semantic density effects in English

4.4.1 Method

As above (§3.4), we obtain LDRT data from the Semantic Priming Project. Since the SPP provides data from a lexical decision *priming* task, as an independent variable we compute the semantic neighbourhood density of the prime word (Buchanan et al., 2001; Shaoul & Westbury, 2010; Siakaluk et al., 2003) by averaging the similarity to the ten closest words. The choice of SOA, dependent variable (z-scores instead of raw RTS), baseline model and data splits followed the previous experiment (see §3.4.1).

For the association norms we cannot use the matrix used above to compute the neighbourhood density of the words. The reason for this problem is that in the SPP the target words are the cues in the Association Norms and the primes are simply responses to the targets. It is not, therefore, necessary that the responses were also given as cues so as to form associative representations. In other words, the word *abdomen* was used as prime to *body*

because given *body* in the association norms some participants responded *abdomen*. However, the word *abdomen* was never given to the participants to list cues. This inconsistency cannot be mitigated by dropping the words which do not have a representation or by imputing values as about two thirds of the dataset primes were not used as targets. To counter this issue, however, we perform a simple solution; recall that we use SVD to reduce the dimensionality of the associative representations. In §2.2.1, we noted that SVD decomposes the original matrix **M** into three matrices **U**, Σ and **V** which when combined give the best approximation of **M**. So far, in order to extract word representations, we have been focusing on **U** which, essentially, retains information about the rows of **M**. However, **V** holds information about the columns of **M**, which in the present case are the responses given to the cues. Theoretically, the information contained in the columns is not the same as the rows contain information about which words were given as responses to a particular cue, whereas the columns contain information about *to* which words a particular word was given as a response. However, this is the best approximation we can think to impute the values for the primes.

4.4.2 Results and discussion

Before looking at the neighbourhood density effects we validate our dataset by looking at whether the similarity estimates from Study 1 are still predictive in the present context. Table 4.2 shows the model fits on the validation and testing sets as outlined above, the improvement of the experimental model over the baseline and Bayes Factors of the experimental models compared to the baseline. As expected from similar studies (Ettinger & Linzen, 2016; Mandera et al., 2017) the lexical decision reaction time task was 'easier' for the models as it resulted in higher gains in all DSMs. Despite the higher fits, the general patterns from §3.4 are preserved in this context; the testing set shows consistently higher estimates than the validation, HAL, LSA and the neural embeddings are the highest scoring models, followed by BEAGLE, COALS and Random Indexing, leaving the two Association Norms with lower fits.

We now turn to the semantic neighbourhood density effects. Contrary to previously reported results (Buchanan et al., 2001; Shaoul & Westbury, 2010), none of the DSMs improved the model fit.[10] Table 4.3 shows the correlations between the SND measure and the reaction times. Regarding simple correlations, only the neural embeddings and the *syntagmatic* models show significant effects. This behaviour is in contrast to the above results, where *paradigmatic* models fared better than the syntagmatic ones. Considering our discussion of mediated priming effects at the end of the last chapter, this should not be surprising. The fact that *syntagmatic* representations encounter, by definition, *more* words than the paradigmatic

[10]Some significant improvements were found for LSA and the neural embeddings, however, the improvement was less than 0.01.

Table 4.2 Summary of the best performing models on the validation and testing sets of the SPP. The predictor variables for each of these models were the same as the baseline model (see text) with the addition of the similarity estimates of the corresponding DSM. B_{m0} shows the Bayes Factor with default mixture-of-variance priors comparing the model that includes the semantic similarity covariate against the baseline. The significance levels refer to the difference between these models and the baseline (see text).

Model	Validation set			Testing set		
	R^2	ΔR^2	B_{m0}	R^2	ΔR^2	B_{m0}
AN_U	0.29	0.02^{***}	4.56×10^{6}	0.32	0.03^{***}	3.64×10^{6}
AN_W	0.29	0.02^{***}	8.22×10^{5}	0.31	0.02^{***}	3.95×10^{3}
BEAGLE	0.3	0.03^{***}	1.71×10^{9}	0.33	0.04^{***}	5.06×10^{9}
COALS	0.29	0.02^{***}	3.26×10^{6}	0.31	0.02^{***}	8.34×10^{4}
HAL	0.3	0.03^{***}	8.31×10^{10}	0.33	0.04^{***}	3.51×10^{12}
LSA	0.32	0.05^{***}	1.11×10^{16}	0.32	0.03^{***}	4.75×10^{8}
Neural Embeddings	0.31	0.04^{***}	1.24×10^{15}	0.33	0.04^{***}	5.86×10^{10}
Random Indexing	0.3	0.03^{***}	1.54×10^{11}	0.32	0.03^{***}	9.62×10^{7}
WordNet	0.29	*n.s*	2.34×10^{1}	0.29	*n.s.*	2.32×10^{-1}

Significance levels: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Note: AN_U = Association Norms (Unweighted); AN_W = Association Norms (Weighted); BEAGLE = Bound Encoding of the Aggregate Language Environment; COALS = Correlated Occurence Analogue to Lexical Semantics; HAL = Hyperspace Analog to Language; LSA = Latent Semantic Analysis; WordNet estimates were obtained using the similarity measure proposed by Resnik (1995) (see §3.3.2).

models give more information about the position of each word in the high-dimensional space. Paradigmatic models, on the other hand, only 'see' sparse information about the possible neighbours for each word rendering them worse predictors. We also compute the partial correlation between the SND measure and the reaction times *given* all the other predictors. In this case, only the neural embeddings show a significant, albeit quite small, correlation, $r(3320) = -0.04$, $p = 0.02$.

The unweighted association norms show interesting correlational patterns. Concretely, both in the simple and the partial correlations, the AN predict an *inhibitory* effect of the SND. These apparent inconsistencies were examined by Mirman & Magnuson (2008) who tested the effect of SND using different feature models, association norms and COALS and different definitions of SND. It was found that words with many near neighbours resulted in longer reaction times, whereas they show an opposite effect for words with many distant neighbours.

Table 4.3 Pearson correlation coefficients and partial correlation between the semantic neighbourhood density measure of each DSM and the standardised reaction times in the lexical decision task of the SPP. The partial correlations were computed as the correlation of the SND measure given all the other variables in the dataset.

Model	Correlation	Partial Correlation
AN_U	0.09^{***}	0.05^{**}
AN_W	0	−0.03
BEAGLE	-0.04^{*}	−0.02
COALS	-0.1^{***}	−0.03
HAL	0	−0.02
LSA	-0.08^{***}	−0.03
Neural Embeddings	-0.11^{***}	-0.04^{*}
Random Indexing	−0.02	−0.02

Significance levels: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$
All $df = 3320$

While our metrics cannot account for this level of granularity, it might still be the case that the neighbours in the association norms are 'nearer' in some sense than those of the DSMs. The reason for this is that in the ANs each word is associated with fewer words (on average, 14.38 words) than in the DSMs (where each word is associated with the entire vocabulary).

Turning to the results from the British Lexicon Project, Table 4.4 shows the performance of the models on the entire dataset. Since we did not tune the parameters on this dataset, from the point of view of the DSMs this dataset can be considered a novel test set, similar to the test set in the SPP. The improvement (ΔR^2) over the baseline is evident for all DSMs. The high fit obtained is in line with previous research; Shaoul & Westbury (2010), for example, report a correlation of 0.42 between the SND estimates of an optimised HAL model and LDRTs from a semantic decision task. While all the models predict significant improvement the estimates of the neural embeddings are the highest, with a difference between that and the next best fitting model of about (0.1) providing without a doubt the best fitting model. All the models predict a facilitatory effect of the SND on lexical decisions, in line with Buchanan et al. (2001) and Siakaluk et al. (2003) and contrary to Shaoul & Westbury (2010). This apparent contradiction can be explained in a similar way as the behaviour of the association norms above.

The unusually high results of the DSMs in this particular task together with the relatively small reduction in the correlations when we partial out the other variables are worrying in the sense that the SND measure might reduce to measuring the effect of other variables. If the

Table 4.4 Summary of the best performing models on the British Lexicon Project. The predictor variables for each of these models were the same as the baseline model (see text) with the addition of the semantic neighbourhood density estimates of the corresponding DSM. B_{m0} shows the Bayes Factor with default mixture-of-variance priors comparing the model that includes the SND covariate against the baseline. *Correlation* and *Partial Correlation* were computed using Pearson's product-moment coefficient (df = 25292). The significance levels refer to the difference between these models and the baseline (see text).

	R^2	ΔR^2	B_{m0}	Correlation	Partial Correlation
BEAGLE	0.3	0.11^{***}	4.05×10^{837}	-0.42^{***}	-0.37^{***}
COALS	0.27	0.08^{***}	9.14×10^{532}	-0.42^{***}	-0.31^{***}
HAL	0.3	0.11^{***}	1.45×10^{819}	-0.41^{***}	-0.37^{***}
LSA	0.22	0.03^{***}	1.98×10^{196}	-0.21^{***}	-0.17^{***}
Neural Embeddings	0.42	0.23^{***}	3.23×10^{2043}	-0.61^{***}	-0.53^{***}
Random Indexing	0.32	0.13^{***}	4.05×10^{815}	-0.41^{***}	-0.35^{***}

Significance levels: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

DSMs are just putting high-frequency words together then we might run into the problem that what the SND really measures is frequency instead of *semantic neighbourhood density*. To this end, we explore the correlations and Variance Inflation Factor (VIF) values (testing for multicollinearity) between the variables of the best performing model (i.e., the neural embeddings). Interestingly, the SND measure correlated significantly, albeit to a very low extent, with the frequency of the word in the BNC ($r(28362) = -0.27$, $p \approx 0$).[11] The rest of the correlations were medium ($|r| < 0.3$) not signalling any multicollinearity problems. The VIF values further support this notion as the value for SND was quite close to 1 ($\text{VIF}_{\text{SND}} = 1.22$), much lower than the common thresholds of 4 and 10 indicating that there were no multicollinearity problems between SND and any other variable.

The high results in the BLP raise some issues on why the DSMs were unable to capture any effects in the SPP. Firstly, one obvious difference is that the SPP is a much smaller dataset than the BLP, which due to the data splitting procedures becomes even smaller. Concretely, the *validation* and *testing* sets combined are about 1/25 of the size of the BLP. While this is important to keep in mind, both the low correlations and the fact that similar effects have been reported on small datasets, render this explanation rather weak. Secondly, the results of the SPP were obtained from American English speakers, whereas the DSMs were trained on a BNC. Taking this dissociation in consideration together with the fact that participants were

[11] We obtain this value after removing 72 outliers with frequency 2.5 standard deviations above the mean. The correlation in the full dataset was $r(28434) = -0.04$, $p \approx 0$.

supposed to be British English speakers[12] might explain the high results in the BLP. While this explanation bears some consideration, the high estimates for the naming task in §3.4.3 do not support this argument. Finally, following the discussion above (§3.6.1), *reliability* should also be an issue here. Indeed, looking at the reaction times between the two datasets (SPP and BLP) on the overlapping set of target words[13] their correlation was quite low ($r = .1$, $p = 0.05$), albeit significant. This corroborates the claims made in Heyman et al. (2018) that item-level generalisation is hard in such tasks.

Looking at the task specifications might elucidate the performance of the models better. Firstly, the two datasets were obtained using different tasks; the SPP was a semantic priming task, where we measure the effect of the *prime* word, whereas the BLP is a simple lexical decision task. While an effect of SND was found in Shaoul & Westbury (2010) in a similar setting as the SPP, we conjecture that this effect will be lower in a priming setting. The reason for this minimisation is that in the priming setting the reaction to the target word would be a function of more variables than in the simple lexical decision. In other words, all things being equal, the effect of SND in a semantic priming task is expected to be lower.

A second –related– argument concerns the display time of the word for which we obtain the SND estimates. Concretely, in the SPP, participants see the prime for 50ms, whereas in the BLP the stimuli remained on screen until the participants gave a response. We argue above (§3.2) that such a low threshold in the SPP is necessary to capture more 'automatic' responses in semantic priming tasks. However, in the case of neighbourhood effects, this low threshold might mask the effect other words have on the activation of the target. The reason for this behaviour is that to compute the SND of a word, one needs to compute its distance to its activated neighbours which presumably requires more time. This notion is further supported when we consider that in Shaoul & Westbury (2010), where similar effects were reported, the display time for the prime word was 500ms presumably leaving time for the co-activation of the semantic neighbours.

Despite the lower scores on the SPP, the results from the British Lexicon Project suggest that our approach of using semantic neighbourhood density as estimated from DSMs to predict lexical decision effects in English is quite successful. The next study explores whether these effects are retained in languages other than English. If, for whatever reason, DSMs can capture such behavioural effects only in English, then, while interesting, cannot be readily used in the SIL simulations reported later (Chapter 5). The potential inability would not only cast doubts on the universality of the effects (discussed further below) but would also pose practical

[12]While participants were recruited from the Royal Holloway, University of London, no special care to select only native British English speakers is reported.

[13]Since each target appears with two primes in the SPP, we report here only the fastest response. However, selecting either prime made no difference in the correlation coefficients ($\max \Delta r \approx 0.01$).

problems in the simulations of the behaviour of speakers of Chinese (§§ 5.4, 5.6 and 5.7). If, on the other hand, such effects are preserved in other languages, then DSMs can capture more 'universal' properties of the human semantic space exploiting language-specific distributional cues instead of 'fitting' the distributional properties of English on other languages. For the next study, we will focus solely on the neural embeddings instead of comparing all the above models. The neural embeddings have proved to be so far the best estimate of semantic representations under various conditions, and by focusing solely on this model, we will reduce significantly the total amount of models needed to train.

4.5 Study 4: Semantic density effects in other languages

4.5.1 Method

We split each baseline dataset into ten consecutive (non-overlapping) folds. We use each fold as our unbiased testing set, while we train a regression model on the other nine. We perform this cross-validation procedure since there are no prime–target pairs in this study (hence no semantic relations to split). At the end of this procedure, we use the average mean squared error of all the held out sets as the loss estimate passed to the Bayesian Optimiser. Subsequently, once we have found the best scoring DSMs, we retrain the linear model using ordinary least squares regression on the entire dataset and compare *that* model to the baseline.

4.5.2 Results and discussion

Table 4.6 displays the fit (R^2), the difference in fit from the baseline (ΔR^2), and the Bayes Factor (B_{m0}) comparing the probability of the data under the model containing the DSM estimates versus the baseline for each language on the held-out sets. Aside from the model fit, Table 4.5 shows the best parameter sets per language. We assess the significance of each parameter by building a separate linear regression model that uses the parameter sets proposed by the Bayesian Optimiser as predictor variables and the average score of the held out sets as the dependent.

Regarding model fit, neighbourhood density was a significant factor in all four languages (Dutch: $F(1, 3213) = 208.96, p \approx 0$, Chinese: $F(1, 5097) = 63.19, p \approx 0$, French: $F(1, 14494) = 551.72, p \approx 0$, Malay: $F(1, 1506) = 26.08, p < 0.001$) indicating that despite their typological differences and associated practical issues (see above), the distributional semantics models can capture behavioural effects in languages other than English. We observe the most substantial improvement in Dutch and French with lower Bayes Factors for Malay and Chinese. However, the size of the corpora in Dutch and French was comparable ($\approx 10^7$

Table 4.5 Best parameter sets for each language reported in Table 4.6 found using the Bayesian Optimiser. The significance of each parameter was assessed by a separate linear regression model which used the results from all the steps the Bayesian Optimiser had to take and predicted the score for the held-out set.

Parameter	Language				
	Chinese	Dutch	English[1]	French	Malay
Dimensionality	50	600	300	600	50^{*}
Window size	$18^{\dagger}$	$2^{\dagger}$	6	18	20^{**}
Subsampling threshold	0	10^{-5}	10^{-5}	0	0
Min count	0^{***}	0^{*}	0	0	0^{**}

Significance levels: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$
[1] These values were determined in Study 1.

words) and substantially higher than Malay and Chinese ($\approx 10^6$ words). This dissociation points to the direction that the more information the corpus contains, the better fit for the model. This latter point is not trivial as more details from the corpus might be irrelevant to the estimates.

4.6 Discussion for Studies 3 and 4

The results obtained in Studies 3 and 4 show that DSMs are not only able to capture the more general structure of semantic memory but also that their predictions extend to languages other than English. In Study 3, using the best performing models of Study 1, we validated the results obtained elsewhere in the literature (Buchanan et al., 2001; Chen & Mirman, 2012; Mirman & Magnuson, 2008; Shaoul & Westbury, 2010; Siakaluk et al., 2003) that the semantic neighbourhood density as computed by DSMs is a significant predictor of LDRTs. Study 4 elaborates on these results postulating that if such effects are found in English, then they should manifest in other languages too. While due to practical considerations, the coverage of the languages examined was quite limited, we find similar effects for SND in Chinese, Malay, French, and Dutch.

The first striking result of the two studies above is the mismatch between the improvement (ΔR^2) in English and the other languages. Concretely, for all the DSMs the average improvement was 0.115, whereas for Chinese, Dutch, French, and Malay 0.025. While, therefore, the SND measure improves the fit of the models for the other languages, it seems that it might be

Table 4.6 Summary of the best performing neural embeddings models predicting LDRTs from neighbourhood density for each language. The predictor variables for each of these models were the same as the baseline model (see text) with the addition of the neighbhourhood density estimates given by the neural embeddings. B_{m0} shows the Bayes Factor with mixture-of-variance priors comparing the model that includes the semantic similarity covariate against the baseline. The significance levels refer to the difference between these models and the baseline (see text).

Language	R^2	ΔR^2	B_{m0}
Chinese	0.32	$+0.01^{***}$	1.59×10^{12}
Dutch	0.34	$+0.04^{***}$	3.56×10^{42}
French	0.41	$+0.03^{***}$	6.06×10^{115}
Malay	0.56	$+0.02^{***}$	1.67×10^{4}

Significance levels: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

something special about English. There are at least three potential explanations we see for this mismatch in performance before we assert that for some reason these models work better in English. Firstly, the quality of the corpus used was much higher than the rest of the languages. As introduced in §2.4.2, the British National Corpus is a pre-processed, human annotated corpus of both spoken and written English containing high-quality texts from professional authors. Comparing that, on the other hand, with the other languages, we see a stark difference. For French, the corpus was crawled from the web, which does not guarantee high-quality texts. Similarly, the Dutch corpus, which contains freely available subtitles, and the Malay corpus, crawled from Wikipedia, are not necessarily written by professionals. The only 'good' quality corpus was, in fact, the Chinese, which, however, was quite small in size. Secondly, the BLP had the practical advantage that trial-level reaction times were made available, instead of item-level averages. This advantage made it possible for us to filter out errors and potential outliers (following the procedure outlined in §3.4.1) in the data, something that we were not able to do for the other languages. A final issue we note is that the BLP baseline was quite low compared to the rest of the languages (even the similar baseline of the SPP). This issue points to the direction that there might be some errors in the data input of the original dataset, which have cascaded to our variable selection procedure.

Another issue was the size of the corpora used in the Chinese and Malay simulations. Practically, this could potentially be a serious problem for DSMs as the sparsity of information would result in lower quality representations. In short, smaller corpora would not provide sufficient information to the DSMs to generate semantic representations that accurately sum-

marise the contexts in which each word appears. The solution the Bayesian Optimiser finds is to (a) reduce the size of the vectors to minimise the potential noise in the representations and (b) to exploit as much information as possible from the corpus. Regarding (a), we have noted that if we increase the size of the vectors to the point where the corpus cannot provide enough information, then the representations are susceptible to noise lowering their quality (see also, Landauer & Dumais, 1997). As for the second point, looking at Table 4.5 we see that the best Chinese and Malay models do not trim the most/least frequent words in the corpus and the window sizes take the maximum value to exploit as much information from the context as possible.

Let us now take a closer look at the parameters for English, Dutch, and French. The English and Dutch corpora are large enough to provide sufficient information to form high-quality semantic representations. As such, the models discard potentially redundant information as high-frequency words which are usually function words such as determiners, particles, prepositions, or pronouns, which have little semantic content. On the other hand, they retain low-frequency words, as such might be discriminative enough to enrich the representations. Interestingly, the window size differed in these two models possibly reflecting linguistic differences. The size of the resulting semantic vectors also differed between Dutch and English; this is most likely the result of the scope of the two corpora; the Dutch corpus encompasses more topics (different sorts of movies) and is more versatile than the English one (which mostly contains texts gathered from newspapers and novels). This last point is corroborated by the fact that the number of different word *types* was 941046 for the Dutch corpus, whereas only 341056 for the BNC. Even after removal of the high-frequency tokens, each word in Dutch was 'seen' in the context of more different words, something that needs larger vectors to be encoded. Similarly, the French corpus also contains a high number of word types (885945) resulting, again, in a larger vector sizes. Finally, we see that the French model did not discard high-frequency words and had an increased window size. The increased window size could be explained by the presence of high-frequency intervening words as the distance between the target, and any meaningful context words is increased. It is unclear why the Optimiser chose not to discard high-frequency words (although this was not a significant parameter), and we consider this to be an artefact of the corpus used.

4.7 General Discussion for Studies 1–4

The primary goal of the four studies presented above was to find the best approximation of the underlying organisation of concepts in the semantic memory. In other words, we sought to determine how we can represent semantic information in the absence of conscious

activation of semantic knowledge. Considering that there are multiple ways in which we can represent semantic information, we test several different kinds of models, each of which makes different predictions concerning the structure of semantic memory. The three primary classes of models we looked at consisted of (a) Association norms as obtained from humans, (b) a large semantic database (WordNet), and (c) word representations formed by exploiting the statistical information inherent in the linguistic environment. Through four studies which look at the reaction times in different semantic tasks, we find that projecting contextual information about each word in a high-dimensional space accounts for the patterns in the behavioural datasets. Concretely, in these four studies we look at the reaction times in tasks of semantic priming directly (Study 1), reported priming effects from studies of mediated priming (Study 2), the effect of semantic neighbourhood density (as measured in these models) in lexical decision reaction times in English (Study 3) and in Chinese, Dutch, French, and Malay (Study 4).

Apart from the main result, another important outcome of the above is the inability of dense representations containing concept-concept relations to capture semantic priming effects. There are a few reasons why this might be the case. Firstly, semantic priming effects without some degree of association between the two words are quite hard to obtain (Lucas, 2000, for a thorough review of the topic). In this vein, it is not surprising that WordNet fails to provide a significant predictor in the regression models. Secondly, we need to take a look at what information these representations contain. We have already argued that one way or the other, DSM representations carry information about the distribution of each word in linguistic usage. On the other hand, WordNet contains rich information about the target concept's locus in a domain independent of language. Retrieving, however, that sort of information from memory in a speeded naming task is a more laborious process than exploiting the surface statistical regularities provided by the language. On the other hand, we see that WordNet fits improve significantly in the longer SOA condition. This interaction points to the direction that such 'deep' semantic information might require some more time to retrieve than surface level regularities (see also, Till et al., 1988, for a thorough exploration of the time course of semantic priming).

However, why do the neural embeddings work better in these tasks? On one level, we argue that this happens because they can model both *syntagmatic* and *paradigmatic* relations (§§ 2.3.1 and 3.6). This property places them closer to BEAGLE, which is the only model that combines both sources of information, albeit performing worse than the embeddings. We argue that this mismatch in performance is a result of the *curse of dimensionality*, from which count models (as BEAGLE) suffer, and which predictive models trivialise. In short, the problem relates to how the model assigns probabilities to novel contexts. By definition,

for *any* DSM the ideal representation $\hat{\mathbf{w}}$ will be one that maximises the following quantity $\hat{\mathbf{w}} = \arg\max_\theta \sum_{j \in C_i} P(w_i | c_j, \theta)$, where w_i is the target word, C_i the contexts in which it can be encountered, and θ the model parameters. The non-neural models estimate this probability by counting the occurrences of w_i in all contexts of C_i, whether C_i is defined as the documents in which w_i appears or as a window around every occurrence of w_i. The problem with this approach is that there exist contexts in which the occurrence of a word would be legal, although not 'seen' in a particular corpus. A consequence of this is that the representations formed by these models will be further away from $\hat{\mathbf{w}}$ as they would not be able to account for this.

Neural embeddings, on the other hand, surpass this problem by considering $P(w|c)$ a continuous probability distribution instead of a discrete one. The consequence of this is that even if w has not been encountered in a particular c, the neural network would be able to impute a non-zero value, instead of directly discarding c. For example, consider that we encounter the word *brown* in the corpus in the following context; 'the quick *brown* fox'. To estimate the representation for the word brown, the model would have to learn the parameters that maximise the following dot products; $\sigma(e_{\text{brown}} \cdot c_j) \forall j \in \{\text{the}, \text{quick}, \text{fox}\}$. Since the vectors at this stage are not normalised, then each dot product exists in the interval $(-\infty, \infty)$, hence, passing this value through the sigmoid function would give either 1 (if c_j is a true context-word) or 0 (if c_j is just noise) (see, §2.3.1, for more).[14] Consider now that the model encounters the unseen sentence 'The quick brown wolf'; models based on counting would assign zero probability to this context. However, the neural embeddings would still consider this a high probability context for the word 'brown' as the only thing that is different is the dot product between $e_{\text{brown}} \cdot c_{\text{wolf}}$. The advantage of this approach is that if $\cos c_{\text{wolf}}, c_{\text{fox}} \approx 1$ (i.e., if the words 'wolf' and 'fox' are closely related), then the model can go beyond the given contexts of *brown* and generalise to 'unseen' contexts.

The performance of the neural embeddings on the task is also encouraging in the context of the discussion in §1.3. We argued there that if connectionism provides an account of human learning processes, then neural embeddings might be relevant to how the semantic system is 'bootstrapped'. Such neural networks learn semantic-like information only by exploiting statistical patterns in the input using simple (and general) mechanisms. In this way, the 'semantic' knowledge we are interested in here is nothing more than knowledge of the statistical regularities in the environment learnt during language processing. That is not, of course, to say that other sorts of semantic knowledge are not present in the human semantic system. This latter point is why we use the word 'bootstrap' above instead of saying that this is how the semantic system is structured. In §7.2, we explore this notion further, testing how this tacit knowledge of statistics can be used to learn 'deeper' semantic relations.

[14] Although this quantity is practically bound by the initial weights and the dimensionality of the vectors.

We argue in §1.3.1 that there are numerous ways of representing meaning, and finding the most appropriate description of that has been called the 'holy grail' of a variety of scientific disciplines (Jackendoff, 2002; Kiela, Bulat, Verő & Clark, 2016). The studies presented above suggest that the distributional patterns of words in the language can account for the priming effects obtained in behavioural experiments. However, such *semantic* distributional patterns do not exist solely in linguistic contexts. Perceptual information from visual (Berzak, Barbu, Harari, Katz & Ullman, 2015), auditory (Kiela & Clark, 2015), and olfactory (Kiela, Bulat & Clark, 2015a) routes contribute significantly to create richer semantic representations. Perhaps, therefore, a complete account of the representations used in these studies would be a fusion of the possible routes that contribute to meaning (see, e.g., the present Shallice, 1988).

The last question that remains from the present chapter regards the value of these representations (i.e., what are they 'good' for). After all, as we have remarked numerous times so far, for any task, the optimal representations will have to be 'tuned' to that task to capture any nuisances in the data. Following this argument, the above exploration can give us –at best– the optimal representations for tasks of semantic priming, which is not what we set out to explore in the beginning. In §3.1, we note that there is a close connection between tasks of semantic priming and those of semantic implicit learning. In both cases, participants have to make semantic decisions without consciously activating their semantic knowledge (or, at least, consider it relevant for the task). From this, we argued that whatever influence semantic knowledge has on these tasks must be exerted from the distribution of concepts in the semantic space rather than ad-hoc conscious categorisations. Chapter 5 explores this notion further, using these representations as input to computational models that simulate the behaviour of participants in SIL tasks.

Chapter 5

Distributional Semantics Approach to Implicit Language Learning

5.1 Introduction

The results obtained in Chapters 3 and 4 present the encouraging view that the distributional patterns of words capture elements of the organisation of concepts in the semantic space. In §3.2, we argue why semantic priming effects give a better view of how human semantic memory is organised. We reach the above conclusion by comparing the predictions of DSMs against human performance in tasks of semantic priming and lexical decision. We then generalise these results by looking at languages beyond English, namely, Chinese, Dutch, French, and Malay, by looking at whether the same predictions hold there as well. Furthermore, in §3.1 we argue why the representations used in modelling tasks of semantic priming are relevant in modelling tasks of semantic implicit learning. In this section, we extend the above findings arguing that the distributional patterns of words can also predict what can be learned implicitly in the tasks presented in §1.5. Also, since DSMs can capture elements of the semantic space of speakers of different languages, we also examine datasets of SIL tasks using English and Chinese as the main language.

5.2 Computational overview of the tasks

In the semantic implicit learning tasks reviewed in §1.5, participants are introduced to a set of novel determiners (Chen et al., 2011; Leung & Williams, 2012, 2014; Williams, 2005) or verbs (Paciorek & Williams, 2015). As noted above, both the presence of the novel word and its

Parts of this Chapter have also been published in Alikaniotis & Williams (2015)

function are made clear to the participants at the start of the experiment. For example, in Leung & Williams (2014) the participants were told that the novel words acted as determiners that encoded distance and at each trial were asked to press a button indicating the animacy of the noun and the meaning of the determiner (see Fig. 1.6). At this stage, therefore, the participants were aware of all the experimental variables except for the co-occurrence rule governing the distribution of this novel word. To provide a computational account of semantic implicit learning, we need to define the *computational problem* that the participants solve during the experiment in the sense of Marr (1982) (see §1.2). Once this is specified, we will move on to specifying the algorithmic details of this problem which result in participants acquiring the co-occurrence rule implicitly.

Formal linguistic analysis aids at abstracting the relevant elements from the experimental stimuli. Despite the intuitive similarity between their structure, the stimuli used are not equivalent between experiments. Linguistically, in the stimuli used by Williams (2005), the critical phrase is a DP where the HEAD (i.e., the determiner) agrees with some semantic feature of the head of its COMPLEMENT (i.e., the noun). Similar constructions were also used by Chen et al. (2011); Leung & Williams (2012, 2014), albeit not always providing sentential context (e.g., Leung & Williams, 2014). Paciorek & Williams (2015), on the other hand, uses a VP as the critical phrase, where the agreement is –again– between the head of the phrase and the head of the NP-complement. Although therefore, the phrases are not equivalent at the outset, there is an underlying connection between them; in all the experiments the agreement is between the head of a phrase (either verb or determiner) and the head of the complement. This analysis not only sidesteps the problem of using different grammatical categories for the first word (determiner vs. verb) but also the issue of an intervening determiner in Paciorek & Williams (2015), where some sentences have the form 'gouble *the* force'. Other problems related to formal linguistic analysis concerning the nature of the features the agreement is dependent upon (Bickerton, 1984; Chomsky, 1995; Kerstens, 1993) are discussed in §5.8.

Under this account, the task of the participant reduces to associating w_c (the word in the complement position) with w_h (the head of the phrase). The experimental procedure employed in these tasks further warrants this simplification as participants are explicitly asked to focus on the critical phrase by repeating it (Paciorek & Williams, 2015; Williams, 2005) while in others they have to make a decision on the current input (e.g., Leung & Williams, 2014). Whatever the cover task is, we assume that in all the experiments, participants solve the *same* underlying computational problem. In other words, given a phrase such as 'gi book' either presented by itself or embedded in a sentence the participant will perform the same computation to associate the two forms. Given the discussions in §§ 1.2 and 1.3, we argue that implicit learning in these tasks arises from the details of the algorithm the participants use to

efficiently processing incoming input. Considering the structure of the phrases, we see two possible computational problems the participants could be solving to alleviate some of the computational cost associated with parsing this input (a) *prediction* or (b) *retrodiction*. In the first case, participants maximise the probability of seeing a particular noun having seen the determiner[1] (e.g., 'lion' given 'gi'). In the latter case, participants view the determiner as a class to which the nouns belong. In other words, in this setting, *lion* would be a *gi* word, whereas *table* a *ro*.

Prediction is the simplest of the two proposals given the order the words appear in the phrases. In short, through exposure to the grammatical system, the participants become sensitive to the DT → N regularity, learning to anticipate N having seen DT. Concretely, this anticipation can be formalised as maximisation of the probability of the upcoming N given the previous words in the sentence (5.6),

$$\arg\max_{w \in \mathcal{V}} P(w \,|\, w_{t-1}, w_{t-2}, \ldots, w_1) \tag{5.1}$$

where $w_{t-1}, w_{t-2}, \ldots, w_1$ is the sentential context. Since the experimental design of the above experiments draws the attention of participants to the relevant phrase, we can simplify (5.1) to include only the relevant elements. This simplification reduces to learning the *forward* (bigram) probabilities, as in (5.2).

$$P(w \,|\, w_{t-1}, w_{t-2}, \ldots, w_1) \approx P(w_t \,|\, w_{t-1}) = \frac{C(w_{t-1}, w_t)}{C(w_{t-1})} \tag{5.2}$$

In this case, the computational problem is similar to the ones employed in standard tasks of statistical learning where participants become familiar with the conditional probabilities governing the distribution of syllables in a linguistic stream (Aslin et al., 1998; Saffran et al., 1996). The main difference between the present task and tasks of statistical learning is that the conditional probability is not contingent on some surface rule (i.e., transitioning from syllable A to syllable B) but on some 'deeper' regularity which associates word A to some feature of word B. The *prediction* proposal is well grounded on a vast array of behavioural, neurological and computational studies (DeLong, Urbach & Kutas, 2005; Dikker & Pylkkänen, 2013; Federmeier, 2007; Frank, Otten, Galli & Vigliocco, 2013; Laszlo & Federmeier, 2009; Lau, Weber, Gramfort, Hämäläinen & Kuperberg, 2014; Levy, 2008, 2011; van Berkum, Brown, Zwitserlood, Kooijman & Hagoort, 2005; Wicha, Moreno & Kutas, 2004) and offers a simple framework to model the above tasks. Despite the abundance of empirical evidence, framing

[1]For exemplification we will refer to w_h as the *determiner* as any arguments apply to Paciorek & Williams (2015) without loss of generality. However, we retain the distinction in §5.5 where we discuss the differences between the experiments.

the computational problem as *prediction* is not without problems. Firstly, prediction is quite hard; without having any knowledge of the underlying system predicting the noun from the determiner is probabilistically difficult as, for the point of view of the participant, *any* noun could follow each determiner. This account is congruent with unreported empirical results (J. N. Williams, personal communication, 2016) using a version of the Leung & Williams (2012) design in which the determiner and noun were presented successively rather than simultaneously. When presented with DET → NOUN constructions (see Fig. 1.6) where the NOUN appears a few hundred milliseconds after the DET, the participants are not faster to make the animacy decision compared to the violation trials. We argue that this drop in performance occurs because for the determiner to be predictive of the upcoming noun, language learners need to either be aware of the rule or prolonged exposure to the system which exceeds the timeframe of the behavioural experiment.

Recent neurophysiological studies comparing the related task of gender processing on L1 speakers and L2 learners, as well as early and late L2 learners show that only L1 and early L2 learners of a gendered language use the determiner predictively (Foucart & Frenck-Mestre, 2010, 2012; Gillon Dowens, Guo, Guo, Barber & Carreiras, 2011; Gillon Dowens, Vergara, Barber & Carreiras, 2010; van Hell & Tokowicz, 2010; Meulman, Wieling, Sprenger, Stowe & Schmid, 2015). This predictive usage of the determiner is indexed by a posterior late positivity (i.e., a P600 ERP) in EEG studies. While the P600 is mostly associated with violations at the stage of syntactic analysis (Hagoort, Brown & Groothusen, 1993), an alternative, but congruent, explanation is that the P600 shows sensitivity to the probabilistic structure of the sentence as low probability *n*-grams are more likely to be ungrammatical and vice versa (Coulson, King & Kutas, 1998; Wicha et al., 2004). Under this account, the fact that only native or near-native speakers of a language show this effect point to the direction that it is their prolonged exposure to language that generates this anticipatory effects (Foucart & Frenck-Mestre, 2012). Apart from neuropsychological studies, experiments using a visual-world paradigm (Lew-Williams & Fernald, 2010) corroborate the above results. In these experiments, participants view familiar objects with names of either the same or different grammatical gender while listening to sentences referring to an object. Lew-Williams & Fernald (2010) found that L1 children and adults orient to the target quicker than L2 learners, when the article is informative about the identity of the noun.

More related to the present tasks, Batterink et al. (2014) using a design similar to Leung & Williams (2012, 2014) tested the performance of participants on SIL tasks recording electrophysiological data during the experiment. After a 90-min nap, participants who were aware of the underlying distributional rule showed a P600 effect, whereas unaware participants showed an early negativity (an effect similar to the N400 ERP but with different spatial distribution).

The P600 effect of gender congruence is similar to what was found in previous studies of L1 and early L2 speakers who use the article predictively. We argue that this dissociation arises because, for predictive effects in this context to arise, language learners, as well as participants in these tasks, need either prolonged exposure to the grammatical system or to have formed a rule that allows them to predict the upcoming nouns.

An alternative, but related, view that avoids the problem raised by uncertainty, retaining the notion of calculating transitional probabilities is to assume that learners rely on the *backwards probabilities*. As shown by (5.3) this quantity measures how likely it is that w_{t-1} will *precede* w_t. Considering the linear order of the words in a sentence, it might seem as counterintuitive that backwards probabilities could be at all informative. However, an array of recent statistical learning studies, as well as computational modelling, are highly suggestive that such *backwards probabilities* are as informative as *forward*. Studies on infants (Pelucchi, Hay & Saffran, 2009), second language learners (Onnis & Thiessen, 2013), corpus analyses (Swingley, 1999) and influential models of word segmentation and chunking (McCauley & Christiansen, 2011; Perruchet & Desaulty, 2008) make use of backwards probabilities as an additional, psychologically valid, source of information. These results might also explain why recent state-of-the-art models of sequence prediction (Graves et al., 2013) are significantly improved when they are trained in a bi-directional manner (i.e., not only from left to right but also from right to left).

$$P(w_t|w_{t-1}) = \frac{C(w_{t-1}, w_t)}{C(w_t)} \tag{5.3}$$

/noindent where $C(w_{t-1}, w_t)$ is the number of counts for the bigram w_{t-1}, w_t and $C(w_t)$ the number of counts for the w_t unigram. The reason such measure is relevant in the present context becomes clearer once we consider again the problem introduced by the increased entropy in *forward probabilities*. We give a brief account of how the problem is simplified when we cast it as *retrodiction* instead of prediction. There are five singular definite articles[2] in the Italian language and roughly 6.6×10^4 nouns (based on itWaCs). For each article, there are on average 1.65×10^4 possible nouns, but for each noun only one article. Even though an article does not always precede a noun, the backwards probability is more useful source of information than the forward. Indeed a quick exploration in an Italian corpus provides support for this point. Using the itWaCs a freely available corpus of Italian texts we counted whether in an article-noun sequence, knowing the article that comes before or knowing the noun that follows would be a better source of information for the learner. Indeed what we find is that the average *forward probability* from any article to the noun is .0009. Contrary

[2]This is counting *l'* twice, both as masculine and as feminine.

to this, the backwards probability from any noun to the article is 0.37 (see §6.3 for a more rigorous examination). Such results are suggestive that instead of *predicting* the upcoming word, the problem is significantly easier if the learner approaches it as a sort of classification.[3] Treating learning of grammatical gender as a classification task is common in tasks of natural language processing (Cucerzan & Yarowsky, 2003; Nastase & Popescu, 2009).

Under this account, the problem at hand reduces to maximising the probability of assigning a noun to its correct class,

$$\arg\max_{y \in Y} P(y|x) \tag{5.4}$$

where Y is the set of possible classes, y is a specific class and x the noun. For example, in Williams (2005), where $Y = \{\text{gi, ro, ul, ne}\}$ given the sequence 'gi dog', we have:

$$\arg\max_{y \in Y} P(y = \text{gi}|\text{dog}) \tag{5.5}$$

Learning these transitional probabilities will guarantee perfect classification of the training exemplar, but will be insufficient in generalising to novel items. This *overfitting* issue can be avoided in one of two ways; firstly, the learner might gradually 'abstract' features of the stored instances and use those to predict the class. Secondly, the learner is equipped with a mechanism that enables assessing the similarity of the stored exemplars to the novel item. The latter has been proposed in many different ways in categorisation tasks (Nosofsky, 1986) and to explain the performance in implicit learning tasks (Chubala, Johns, Jamieson & Mewhort, 2016). We mainly explore the first proposal, arguing in §5.8.3 why the second one would face problems in the present context. Either way, both proposals assume that the computational model maximises the following probability:

$$\arg\max_{y \in Y} p(y|\mathbf{x}, \theta) = f(\mathbf{x}, \vec{\theta})_y \tag{5.6}$$

where Y is set of determiners/verbs, $\mathbf{x}$ is the semantic representation, θ are the model parameters, and $f(\mathbf{x}, \vec{\theta})$ is the model output as computed in (5.7).

Framing the computational problem this way gives a natural explanation for the generalisation task in these experiments where the participant is forced to choose between two alternatives where the only thing that changes is the determiner/verb:

$$P(y = \text{gi}|\text{Mary patted \_\_ tiger in the zoo})$$

[3] In languages where articles or determiner-like particles *follow* the noun, the forward probabilities should be more informative than the backwards.

$$P(y = \text{ro} | \text{Mary patted \_\_ tiger in the zoo})$$

If the model parameters θ which were set to maximise (5.6) did not use any predictive regions of the input semantic representations, then both hypotheses would be equiprobable. If on the other hand, to be able to carry out the classification, the participants set their 'internal' parameters θ in a way that maximises (5.6) but also promotes generalisation, they effectively learn the system 'implicitly'.

5.3 Method

5.3.1 Feed-forward neural network

Figure 5.1 illustrates the architecture of the feed-forward neural network used in the simulations. The input units are shown on the left and activation propagates from left to right to the output. The size of the input layer is the size of the vocabulary (ca. 120000) and the size of the representation layer is the size of the word embeddings (300). We substitute the input → representation matrix $\mathbf{W}_{ri}$ with the embeddings matrix $\mathbf{M}$ obtained in §3.4 and keep it 'frozen' during training (i.e., weights are not updated). Using this scheme, we present each word to the network as a one-hot vector (e.g., $\langle 1\,0\,0\,0\,0\,\ldots\,0 \rangle^{\top} \in \mathbb{R}^{|\mathcal{V}|}$, where the 1 is the index of the target word in $\mathcal{V}$) the dot product of which with the input → representation matrix yields the neural embedding in the representation layer. Since the representation layer has linear activation, no transformation is applied to the embeddings. The entire set of units for the hidden and output layers is shown in the figure. Since all the behavioural experiments reviewed use a four-class system (either determiners or verbs), the size of the output layer remained the same throughout all the simulations. For every simulation, the network is trained to turn on the units in the output layer, which correspond to the categories (either determiners or verbs) in the behavioural experiment.

Given this architecture, the overall objective of the model is to learn to associate word representations as extracted from large linguistic corpora to novel determiner classes as in the behavioural experiments. For example, a sentence in the Williams (2005) dataset which read '*The fire brigade had to rescue* ***ul cat*** *from the top of the tree.*' would become the input–output pair *cat*–*ul* as depicted in the figure. Since the network has access to the entire vocabulary which we activate given an external probe, we can think of this process as activating a long-term memory trace of this word (e.g., Kintsch & Mangalath, 2011) which is subsequently paired with the novel element. Our main goal in the simulations is to see how the network would behave to unseen words once it has learned to associate pairs from the training sets.

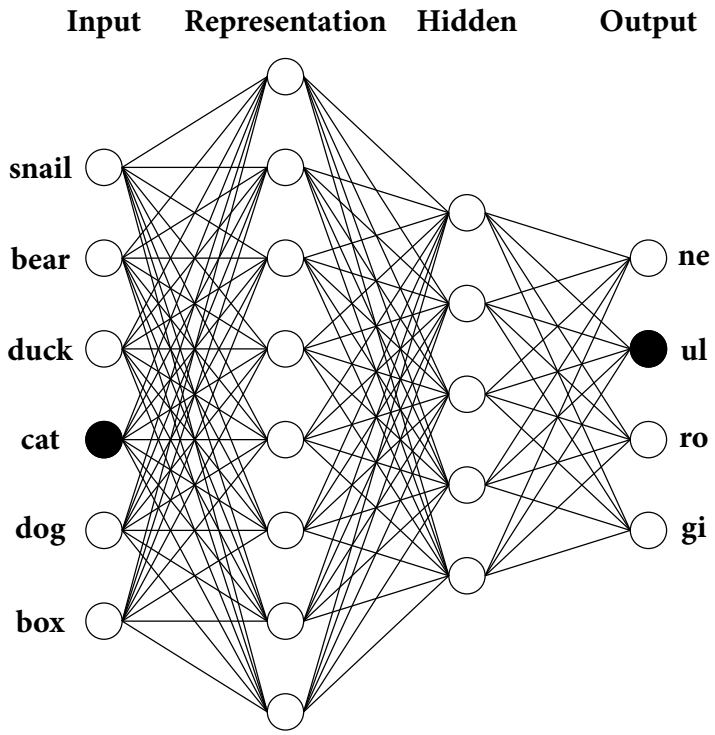


Figure 5.1 Depiction of the connectionist model of classification used in the simulations. Input units are shown on the left and activation propagates from left to right. For illustration purposes we only show a subset of the units used in the input and the representation layers. The size of the input layer is the size of the vocabulary (ca. 120000) and the size of the representation layer is the size of the word embeddings (300). The entire set of units for the hidden and output layers is shown in the figure. Each unit in the input layer corresponds to a word in the corpus and its activation in the representation layer corresponds to the neural embedding as described above. For every simulation, the network is trained to turn on the units in the output layer, which correspond to the classes (either determiners or verbs) in the behavioural experiment.

The function that the network in Fig. 5.1 ends up computing given a semantic representation $\mathbf{x}$ and parameters θ is

$$f(\mathbf{x}, \theta) = \mathcal{S}\big(\mathbf{W}_{oh} \cdot \sigma(\mathbf{W}_{hr}\mathbf{W}_{ri}\mathbf{x} + \mathbf{b}_h) + \mathbf{b}_o\big) \tag{5.7}$$

where σ is a nonlinear function (for the simulations reported here use the hyperbolic tangent function), $\mathcal{S}$ is the softmax function, $\mathbf{W}_{hr}, \mathbf{W}_{oh}, \mathbf{b}_h, \mathbf{b}_o \subset \vec{\theta}$ are the learnable parameters of the model denoting the representation to hidden and hidden to output matrices, as well as the two

bias vectors from the hidden and output layers ($\mathbf{W}_{ri}$ is also a subset of θ but not a trainable matrix).

Initially, the connections between the units (i.e., the matrices $\mathbf{W}_{hr}$ and $\mathbf{W}_{oh}$) have small random values so that no category is preferred *a priori* by the network. The initialisation of those weights is an integral part of the learning procedure as it can be the case that the network is unable to learn the patterns given an improperly initialised weight configuration. We follow the initialisation procedure proposed by Glorot & Bengio (2010), which takes into account the size of each layer in the network (more details in §B.1), and is shown to give better results in multilayer networks. The network learns to perform the task by finding a configuration of weights such that given a semantic representation in the input layer, it activates the node for the correct class in the output layer, inhibiting the activation for the incorrect classes.

Finding the appropriate configuration of weights is not a straightforward process, and many algorithms used in the literature have been criticised in that they are not biologically plausible. Although deriving 'biologically' plausible learning algorithms is an active area of research (Scellier & Bengio, 2017) we train the network with the commonly used **backpropagation** algorithm (Rumelhart, Hinton & Williams, 1986a). As noted in §1.3, backpropagation is an iterative process by which the network makes small adjustments to its weights every time it makes an incorrect prediction. The objective is that the next time the same activation pattern appears in the input, the prediction will be closer to the teaching pattern. Effectively, after a number training cycles (which we call 'epochs') where in each cycle the network sees all the items in the training set in random order, the network will reach a state where given an activation pattern in the input layer it will activate the correct nodes in the output layer.

How we quantify the 'prediction error' the network is making is a major factor in the discussion as it can not only change the results but also our interpretation of them. Intuitively, we want to quantify the difference between what the network predicted for its output and what the output was supposed to be. From a probabilistic perspective, each teaching pattern can be interpreted as a degenerate discrete probability distribution over classes as the correct alternative always has probability 1 while the rest 0. On the other hand, the normalisation factor in the denominator of the output layer's softmax function (5.6) ensures that the network's predictions sum to 1, prompting us to look for a measure of distance between two probability distributions. A commonly used function in information theory which measures this distance is the **cross-entropy error**. Given a true distribution (the teaching pattern) p and a coding distribution (the network's prediction) q, their distance can be quantified as

$$H(p, q) = -\sum_{x} p(x) \log q(x) \tag{5.8}$$

where p is the true distribution, q is the network's prediction and x is a particular example. In other words, the participant during the experiment learns the probability that a certain determiner precedes a noun. For example, given the word 'monkey' and the potential labels 'gi' 'ul' 'ro' 'ne', an output layer of $\langle 0.3\ 0.6\ 0.03\ 0.07 \rangle^\top$ would mean that the most probable label for 'monkey' would be 'ul'. However, the activation of 'gi' is still higher than those of the determiners that co-occur with inanimate nouns rendering it the preferred choice when forced to choose between 'gi' and either 'ro' or 'ne'.

There are three interrelated issues with our training procedure which all stem from the limited amount of training data and the high-dimensionality of the input. Firstly, the number of free parameters is quite large in a neural network. More specifically, the number of parameters in a neural network with one hidden layer is $D \times H + H \times O + H + O$[4] where D is the dimensionality of the input vector, H the size of the hidden layer and O the size of the output layer. As an example, in a neural network where the size of the input vector is 300, the size of the hidden layer is 5, and that of the output layer is 4 the total number of parameters is 1529.[5] We counter this issue by applying a penalty to the model parameters during learning. Concretely, during learning, we add a term in our cost function to prefer smaller weights (commonly called the weight decay) $\lambda/2n \sum_w w^2$, where λ controls the magnitude of the weights. This way the solution learnt by the model penalises larger weights, placing, thus, importance on spurious elements of the input. We experimented with various values for λ and found that, empirically, a value of 0.1 (i.e., preferring to minimise the cost function instead of small weights) worked best regarding minimising the error on the test set.

Secondly, another counterargument would be that the neural embeddings contain a lot of noise which prohibits the network from discovering interesting regions in the input. To counter this problem, we *dropout* (Hinton, Srivastava, Krizhevsky, Sutskever & Salakhutdinov, 2012) weights from the input layer. *Dropout* is a simple technique by which during training some nodes of the matrix are randomly "turned off" (i.e., set to 0). This technique has been widely used in machine learning to avoid overfitting the dataset by focusing on noisy regions. Moreover, *zero*ing elements of the feature matrix has been used extensively in computational modelling of memory processes (Hintzman, 1986) as denoting imperfect recall. In other words, an interpretation of this procedure would be that during the training phase, the participants do not retrieve perfectly the distributional representation from their semantic memory.

Thirdly, because the number of datapoints is quite small, the optimisation algorithm is more prone to local minima. In other words, the network selects a solution that does not

[4]More generally, in any fully-connected feedforward neural network the number of parameters would be $\sum_{i=2}^{L} L^{i-1}L^i + L^i, \quad i \geq 2$ where L is the number of layers in the network and L^i the size of the i-th layer.

[5]Since we do not carry on training the semantic representations, this does not include the first layer of the matrix as in Fig. 5.1.

minimise the error but cannot move away from that because there no other solution in the immediate region that minimises the cost. Although the solutions to the first two problems aid in solving this issue too, we opt for re-running each simulation 30 times, then averaging the results. In this instance, we can consider each run as an independent learner, who might get stuck with a local solution, then by averaging their performance on the test set we effectively perform a by-subjects analysis.

5.3.2 Evaluation

General approach

The output of the network in Fig. 5.1 is a discrete probability distribution, where each element is the probability that the input noun is associated with a particular class. Using Luce's choice axiom (Luce, 1959), modelling the 2AFC tasks is straightforward as the probability of selecting the correct alternative is equal to its probability over the sum of the probabilities of the two alternatives. Concretely,

$$P(c) = \frac{f(\mathbf{x}, \theta)_c}{f(\mathbf{x}, \theta)_c + f(\mathbf{x}, \theta)_i} \tag{5.9}$$

$$P(i) = 1 - P(c) \tag{5.10}$$

where $f(\mathbf{x}, \theta)$ is the model output given input $\mathbf{x}$ and parameters θ, and c, i are the indices of the correct and incorrect alternatives, respectively.

Modelling the reaction time tasks is less straightforward as there is no direct comparison of grammatical vs. ungrammatical. However, following the analyses in the behavioural tasks, we compare the average activation of the control trials to that of the violation trials.

Point estimation

Using the above methods, we can keep track of the generalisation gradients at each epoch throughout training. While this is useful as it enables us to examine the developmental stages of learning, it does not let us see how well it quantitatively fits the data. While the general *trends* in participants' performance may be describable by the gradients, the experiments in question use *point estimates* to describe the performance of the participants. If we are to directly compare the predictions of our model to the behavioural results, we then need a principled way of selecting the epoch which we believe closely matches the behavioural performance. A solution to this issue would be to take the *Maximum Likelihood Estimate* (i.e., the estimate which best fits the data). For example, consider the 2AFC generalisation task in

the second experiment from Williams (2005) where participants achieve a 0.65 generalisation rate. In other words, they select the correct alternative 65% of the time. If the performance of the model at epoch e is 0.65 but at $e + 1$ is 0.70, then it might be desirable to choose e instead of $e + 1$ as it maximises the likelihood of the behavioural data. Concretely, taking the *Maximum Likelihood Estimate* (MLE) is tantamount to choosing the epoch where,

$$\arg\min_{e} \quad L(\overline{\mathbf{y}}, out_e) \tag{5.11}$$

$$out_e = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}_i, \theta)_c^e \tag{5.12}$$

where e is the e-th epoch, L is the loss function as in (5.8), $\overline{\mathbf{y}}$ is the average performance in the task (as given in the original study), and $f(\mathbf{x}_i, \theta)_c^e$ is the output of the network for the correct alternative at epoch e given input $\mathbf{x}_i$. Averaging over these inputs, we get the model's generalisation rate at a given epoch.

5.3.3 Visualising the spaces

One key aspect of data exploration and analysis involves being able to visualise high-dimensional data. However, as humans, we are limited to being able to visualise up to three dimensions.[6] Data visualisation in the present context becomes particularly important as it enables us not only to see how the words are distributed in the semantic space but also to explore how is the network arriving at the solutions by visualising the activation of the hidden layers at certain inputs and epochs. Dimensionality reduction methods commonly used by psychologists attempt to either find the eigenvectors which capture the most variance in the data (e.g., PCA, Hotelling, 1933) or find a spatial configuration in a n-dimensional plane which most closely preserves the dissimilarity matrix (e.g., Multidimensional Scaling (MDS), Torgerson, 1952) see de Oliveira & Levkowitz (2003) for a review.

Here we adopt the **t-Distributed Stochastic Neighbor Embedding** (t-SNE) (van der Maaten & Hinton, 2011) which has been found to work better in the context of word embeddings (Hashimoto, Alvarez-Melis & Jaakkola, 2016). t-SNE constructs a low-dimensional space by preserving the structure of the high-dimensional space. Concretely, consider two high-dimensional data points x_i $x_j \in \mathbb{R}^n$ and let $p(j|i)$ be the conditional probability that x_i would pick x_j as its neighbour. Now consider two random low-dimensional data points y_i and y_j and their conditional probability $q(j|i)$. The objective of t-SNE is to minimise the divergence

[6] This is not entirely true; humans without spatial training are particularly able to reason in four-dimensional tesseracts (Ambinder, Wang, Crowell, Francis & Brinkmann, 2009). Furthermore, we can visualise four dimensions by projecting the fourth dimension as a colour map on a three dimensional manifold. However, for this thesis, this will not be necessary as two-dimensions are enough.

between P and Q by learning appropriate coordinates for y_j and y_i. In all the figures which follow and are the result of a dimensionality reduction technique, we have omitted the x and y scales. We do so because in the final solution the scale and orientation are irrelevant and unnecessary, and meaningless comparisons might disrupt the reader.

5.3.4 A note on non-distributional representations

Before reporting the results of the simulations, we examine whether WordNet (see §1.3.1) could predict the behavioural results in the experiments outlined in §1.5. Following Harm & Seidenberg (2004) and Monaghan, Chang, Welbourne & Brysbaert (2017) we exploit the hierarchical structure of WordNet forming semantic representations using as features the synsets which either a hypernymic, a holonymic, or a meronymic relation to the target.[7] Once we obtain the features for every word, we construct a $|\mathcal{V}| \times |F|$ matrix, where $|F|$ is the size of the feature set ($|F| = 28568$), and, similarly to the association norms, assign a value of 1 to the corresponding elements of the matrix. Since the original dimensionality of the WordNet vectors is quite high for input to the neural network and because we have very few datapoints for each experiment, we explore the use of dimensionality reduction on the WordNet vectors. Using SVD and the similarity datasets discussed in §3.6 we reduce the dimensionality of the original dataset to 1000 elements (see also, §A.5). We finally selected the synsets corresponding to the words used in the experiments by maximising the within group similarity. Concretely, for every semantic group in each experiment (e.g., *animates* in Williams, 2005) we select the alternatives which maximise the average group similarity. While the complexity of this method grows exponentially with the number of words used in each experiment, using a dynamic programming approach substantially reduces the computational cost.

Table 5.1 summarises some statistics for the WordNet representations such as the average number of features (on the non-reduced matrix) as well as the similarity between the test sets and the training set for each group. We note two important results of this table regarding the WordNet representations. Firstly, in most experiments the similarity between the testing items and the training items from the same category is higher than the similarity against the different category, crudely predicting learning effects. Secondly, the number of features for the concrete stimuli in the Paciorek & Williams (2015) experiments is greater than that of the abstract items, agreeing with the proposals from (Plaut & Shallice, 1993) that concrete words are supported by more semantic features than abstract ones.[8]

[7] A complete list of the synset-feature mappings can be found at `https://github.com/dimalik/wordnet_features/`

[8] While for the distributional models the words that we can use are limited to the breadth of the corpus, for WordNet the words . Three collocations could not be found *reply slip*, *sim card* and *screen protector* and were deleted. To avoid unbalanced sets, we also removed at random three items for the opposite group (e.g., if *reply slip* appeared in the FLAT group we removed a word at random from the long group). Furthermore,

Table 5.1 Summary of the simulated semantic implicit learning experiments. The table contains the relevant semantic distinction, the language in which they were administered, the average number of WordNet features, and the within and between groups similarity in either WordNet or DSMs. The average within similarity was calculated as the average of the upper triangle of the square similarity matrix, whereas the between similarity was calculated by firstly averaging the similarity of every word in one group to every word in the other, and then average all of them together.

Experiment	Semantic distinction	Language	WordNet			DSM	
			#Features	Within	Between	Within	Between
Williams (2005)	English	Animate	14.75	0.78	0.39	0.4	0.15
		Inanimate	9.56	0.75	0.42	0.33	0.16
Chen et al. (2011, Exp. 1)	Chinese	Animate				0.44	0.27
		Inanimate				0.28	0.25
Chen et al. (2011, Exp. 3)	Chinese	Large				0.27	0.13
		Small				0.12	0.25
Leung & Williams (2012)	English	Large	15.33	0.85	0.66	0.32	0.23
		Small	13.05	0.68	0.71	0.29	0.29
Leung & Williams (2014)	English	Flat	8	0.34	0.39	0.2	0.2
		Long	10.09	0.36	0.32	0.21	0.2
Leung & Williams (2014)	Chinese	Flat				0.27	0.23
		Long				0.23	0.23
Paciorek & Williams (2015, Exp. 1)	English	Abstract	5.83	0.57	0.23	0.25	0.06
		Concrete	11.7	0.59	0.26	0.39	0.09
Paciorek & Williams (2015, Exp. 4)	English	Abstract	5.62	0.62	0.25	0.23	0.09
		Concrete	8.58	0.47	0.23	0.24	0.09

Note: The WordNet estimates concern only the English stimuli (see text), whereas we compute the DSM estimates for the non-English experiments using the models gathered in §4.5.

We compute WordNet similarities using the cosine distances between the WordNet feature vectors instead of any of the graph distances outlined in §3.3.2.

To our knowledge, there are four attempts at providing a Chinese equivalent to WordNet. The The Sinica Bilingual Ontological WordNet (BOW) (Huang, Chang & Lee, 2004; Huang, Tseng, Tsai & Murphy, 2003), the Southeast University WordNet (SEW) (Xu, Gao, Pan, Qu & Huang, 2008), the Taiwan University WordNet (CWN) (Huang, Hsieh, Hong, Chen, Su, Chen & Huang, 2010) and the Chinese Open WordNet (COW) (Wang & Bond, 2013). While all four derivatives offer a version of WordNet in Chinese, they constitute translations of the original English version instead of a separate project. For example, the SEW uses pattern matching algorithms between the English definitions provided by WordNet and a Chinese-English dictionary to match Chinese words to English synsets. Building on that logic the other versions improve on this method by using better databases to match between or even combine sources. However, at their core, they use the same taxonomy.

WordNet predicts that the systems based on the 'core' semantic distinctions (animacy and concreteness) should be learnable as the similarity between the synsets of the same category is higher than that of the other category. However, WordNet was constructed under the assumption that speakers of different languages share semantic representations, so there should be no difference between speakers of Chinese and English regarding the long–flat distinction. The question now becomes whether a Chinese version of WordNet based on novel data would reflect this discrepancy. We explore this issue further in §5.8.1.

5.4 Study 5: Animate / Inanimate

5.4.1 Introduction

Studies on the categorical organisation of conceptual knowledge suggest that animate and inanimate concepts are represented differently in mind. While the categorical organisation of conceptual knowledge is a contentious issue as most researchers view such semantic distinctions to be the two ends of a continuum (e.g., Brysbaert, Warriner & Kuperman, 2013), the animate-inanimate distinction seems to be well-grounded based on neuropsychological evidence. Caramazza & Shelton (1998) review several studies on selective brain damage (see also Capitani, Laiacona, Mahon & Caramazza, 2003) as well as providing original evidence from a brain-damaged subject who had the inability to name animate objects. After a rigorous examination on a battery of tasks, they conclude that the obtained patterns cannot be explained solely on sensory/functional grounds (e.g., animate objects yield similar sensory responses), giving support to a categorical distinction between animate and inanimate con-

due to British-American English differences, a few words were replaced (e.g., *movie ticket* → *theater ticket*). Lastly, wherever possible, we chose similar words which would possess the same featural representations (e.g., *icepop* → *lollipop*). For more details on the datasets used see Appendix B.

cepts. Caramazza & Shelton (1998) further argue that such distinctions are to be expected for categories for which there are evolutionary survival pressures to be developed separately (Gelman, 1990).[9]

An issue related to the above is whether this suggested distinction between animate and inanimate concepts is manifested in any way in linguistic usage. This implication is relevant to our models as if this distinction is reflected in the distributional patterns of words; then we would expect the DSMs introduced above to be able to capture it. The results presented in Table 5.1 show that neither WordNet nor the neural embeddings have any trouble distinguishing between animate and inanimate concepts. However, while WordNet synsets are marked for animacy as they are all descendants of `living_thing.n.01` (see Fig. 1.4), it does not necessarily mean that the differences in the DSM are a result of different distributional patterns between animate and inanimate concepts. The categorisations usually used in these experiments are too restrictive (animals vs. furniture pieces) giving rise to alternative explanations. In what follows we show some examples from English and Bantu languages showing that animacy often determines word order and morphological marking.

Looking at the Bantu languages, we see that they use different morphological markers to distinguish between animate and inanimate nouns. Examples (5.13) and (5.14) show the distributional patterns in Congo-Swahili in which adjectives, connectives and the verb have to agree with the grammatical marker of the noun which is widely determined by its semantics.

(5.13) *Batoto* *ba-mingi* *b-a* *iki* *ki-pande* *ba-li-kwa-ka-po*
2-children 2-many 2-CON DEM.7 7-piece 2-PST-be-ASP-LOC
'Many children from this area were there.'

(5.14) *Bi-le* *bi-ntu* *bi-ote* *bi-li-kwa-ka* *mw-a* *Kalumbu* *bi-l-ingia* *umu*
8-DEM 8-thing 8-all 8-PST-be-ASP 18-CON Kalumbu 8-PST-enter DEM.18
mu-nyumba
18-9.house
'All the stuff that was in Kalumbu's (house) was transferred into this house.'

where 2 and 8 denote the corresponding noun class (2 = living things; 8 = things). Such constructions in which elements of the sentence agree with a semantically determined noun class are well attested in many languages from distinct language families such as *Dyirbal* (Dixon, 1972) and *Bininj Gun-Wok* (Evans, 2003). In these languages, distributional patterns can signal the presence or absence of an animate or inanimate concept.

Regarding word order, *animacy* distinctions are quite common in the world's languages. From a typological point of view, the presence or absence of an animacy feature is important

[9]According to Caramazza & Shelton (1998), these survival pressures stem from the fact that *animals* are potential predators but also a source of food, whereas *plants* are sources of food and medicine (p. 20).

in determining word order choices in dative alternation and agreement patterns in languages from different language families (Evans, 1997; Hawkinson & Hyman, 1974; Morolong & Hyman, 1977; Polinsky, 1996). In the case of English, while animacy is not marked overtly as in Swahili (see above), Bresnan & Hay (2008) and Hinrichs & Szmrecsányi (2007) present evidence that in English varieties spoken in New Zealand as well as in the US and the UK, animacy is a significant predictor of the choice of syntactic paraphrases such as dative and genitive alternation.

Finally, current syntactic theories (Chomsky, 1995; Gazdar, Klein, Pullum & Sag, 1985) propose that during the computation of the grammar of a sentence, certain semantic features are 'checked'. In other words, the elements of a sentence need to have some semantic congruence which is determined during syntactic parsing. While the necessity of this 'checking' varies between theories as well as between languages, it seems that theories of syntax converge on the fact that some 'core' conceptual distinctions are available to the speaker at all times. In the present context, this becomes necessary as various researchers have argued that a [±ANIMACY] feature exists between genetically diverse language in determining the arguments of a DP (see Adger & Harbour, 2007, for an application in Kiowa and the references therein for applications to other languages).

Considering the above discussion, we expect that the neural embeddings would be able to capture the effects observed in the behavioural results in that the grammatical system would be learnable. However, since for the DSM, *animacy* is only indirectly inferred from the distributional patterns of the words would not expect such high generalisation rates as obtained by WordNet (where animacy is marked as a distinct feature).

5.4.2 Materials

We simulate the performance of the English speakers using the dataset from Williams (2005, Experiment 1), which uses 24 nouns split evenly between animate and inanimate as seen in Fig. 5.2a. Using the embeddings matrix from §3.4, we generate training data by pairing the indices of the columns of the word vectors used in the experiments with one-hot vectors (localist representations) corresponding to the novel determiners. As explained above, given the index of the word in the input layer, the activation of the representation layer (i.e., the dot product between the input and the embeddings matrix) will be equal to that of the neural embedding with the same index formed in Chapter 3. We train and evaluate the performance of the model using the train-test split by Williams (2005) (see §C.1).

For the Chinese speakers, we follow the same procedure as above, substituting the embedding matrix with the Mandarin Chinese one formed in Chapter 4. The experimental stimuli and procedure matched that of Williams (2005) so no additional changes were made.

5.4.3 Results and discussion

We evaluate the performance of our model in this and any subsequent experiments on three grounds. Firstly, whether the generalisation gradients of the network retain the observed patterns in the behavioural data, secondly, whether the predictions of the network for the 'correct' epoch match those reported in the studies, and, finally, how the hidden layer responds to the training input. This last point is of particular importance as it can help us understand the solution that the network has found to classify the training data. For example, the network might achieve high levels of generalisation, without using the intended semantic distinction. Examining the activation of the hidden layer then helps us understand which regions of the input the network considers to be relevant.

Figure 5.2 shows a two-dimensional projection of the stimuli used in the two animacy experiments. Fig. 5.2a plots the distributional vectors of the words used by Williams (2005), whereas Fig. 5.2b plots the same distributional vectors when trained on a Chinese corpus. For illustration purposes, we substitute the Chinese characters with their equivalent English translations. We obtain both distributional matrices from the simulations in Chapters 3 and 4, colour-coded for animacy (i.e., 'green' for animate and 'orange' for inanimate words). Despite some minor inconsistencies in the Chinese embeddings, in both cases, the problem should be relatively easy for the model as the stimuli are *linearly separable* (i.e., one can draw a line that demarcates between the two groups). However, we note at this point, that the *primary* goal of the network is not to distinguish between animate and inanimate concepts but to learn to judge the alternative groupings of already learnt associations. For example, given the configuration of the words in the semantic space, if a determiner were seen with *bear*, *snake*, and *monkey*, would the learner by more inclined to generalise to *bee* or to *book*? While the problem is simplified if the network has knowledge of concrete semantic features, the critical point is to associate those features with the relevant determiners. In other words, even if the network 'understands' the difference between animate and inanimate concepts in general, it does not mean that it will associate those features with the correct determiners.

Turning now to the network's performance on the test sets, Fig. 5.3 plots the performance of the network on the two datasets (the English and the Chinese) both overall and broken down by semantic category (animate vs inanimate). Figure 5.3a plots the overall generalisation performance of the network on the English test set. We observe that the network plateaus after a few dozen epochs at 55% (which is our point estimate) accuracy and then does not improve after that. This accuracy rate is somewhat lower than the one reported in the behavioural study (59%) but still on the same scale. Given the *t*-SNE solutions in Fig. 5.2, this result points to the direction that either the problem of re-associating the already paired determiners is 'unlearnable' or that the stimuli used in these experiments cannot yield perfect generalisation.

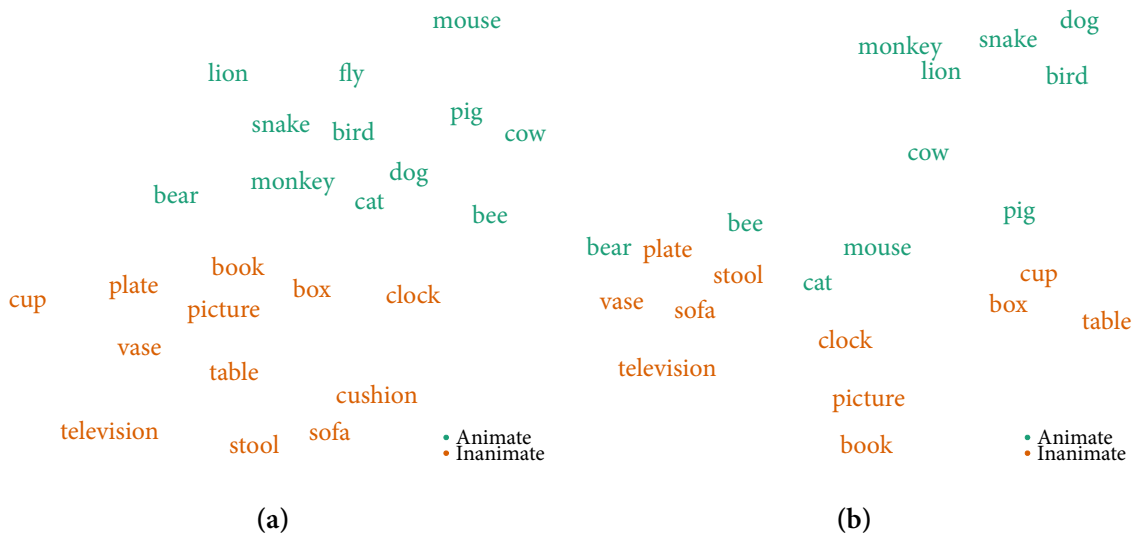
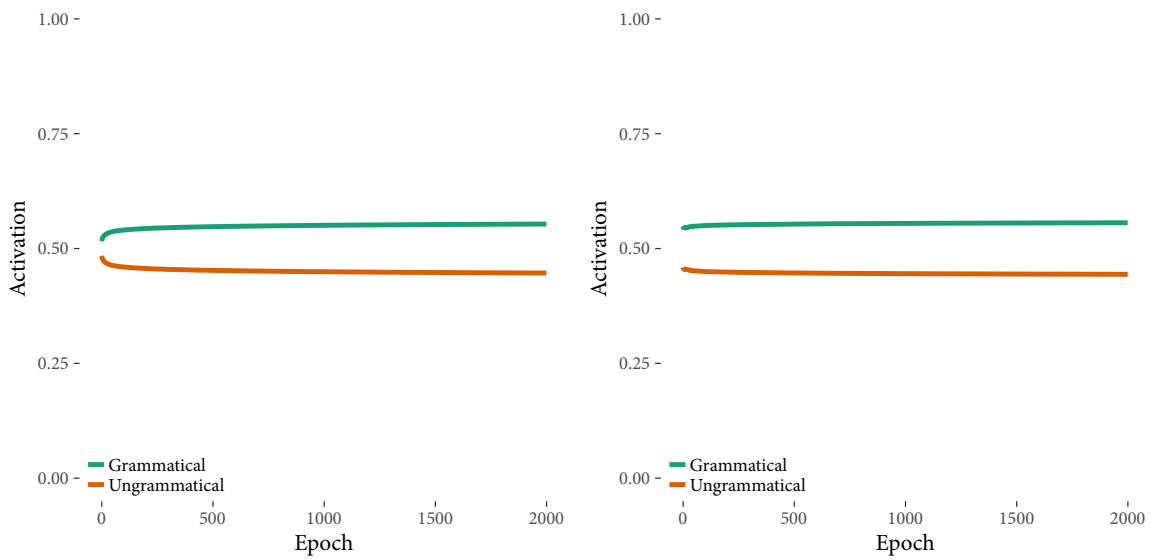


Figure 5.2 Two-dimensional projections of the stimuli used in the animacy experiments. (a) Projection of the words used in the English experiment (Williams, 2005), (b) Projection of the words used in the experiments with speakers of Chinese (Chen et al., 2011, Experiment 1). We translate the words into their English equivalents only for illustration purposes; for more details on the Chinese datasets see Appendix C.

In the first case, the problem is that the network identifies distinct neuronal ensembles in the input for each determiner. This behaviour causes problems during testing as there is no overlap between the words that belong to the same category but were paired with distinct determiners. In the second case, the network discovers the relevant regions but associates them with weak connections to each determiner causing the residual activation during testing to be small. In other words, the network is expressing uncertainty by assigning small weights to the relevant regions of the input to the determiners. When probed with an already paired word, the activation of the other grammatical alternative is higher than the rest but still quite low overall. We attempt to answer these problems in the next paragraph when we evaluate the activation of the hidden layer. For the Chinese speakers, the network exhibits performance similar to English, although the point estimate is now equal to the behavioural data (.56 for the network, .56 ±10 in the behavioural data).[10]

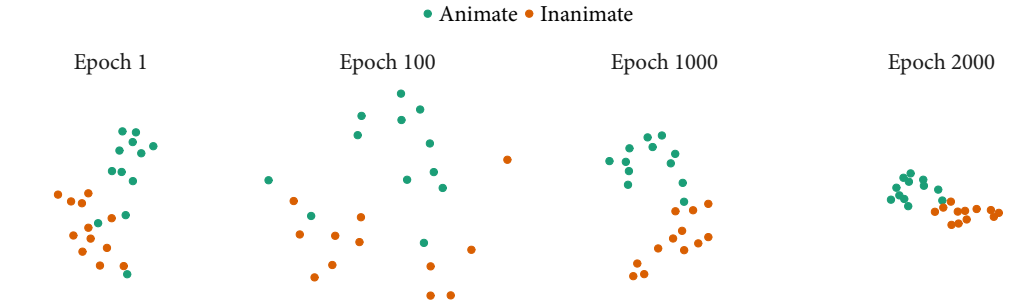
We finally turn to the activation of the hidden layers. We noted above that we could attribute the limited generalisation performance of the network to either the network not being able to discover the regions of interest in the input vector or that these regions are

[10]We obtain this estimate from the responses based on unconscious structural knowledge (Chen et al., 2011, p. 1754).

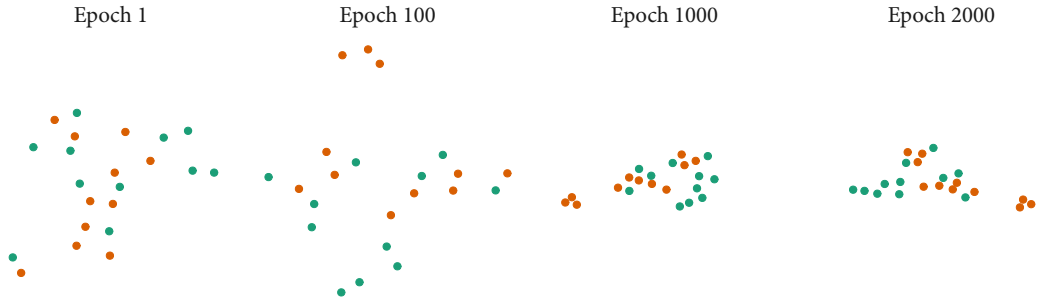


(a) Generalisation performance (English)

(b) Generalisation performance (Chinese)



(c) Activation of the hidden layer (English)



(d) Activation of the hidden layer (Chinese)

Figure 5.3 Generalisation gradients for the two animacy experiments. Fig. 5.3a plots the by-epoch performance of the model on the English dataset, whereas Fig. 5.3b provides a similar view for the Chinese. Figs. 5.3c and 5.3d plot the activation of the hidden layer when probed with the stimuli of the training sets. Concretely, We feed the network all the stimuli without performing any weight updates and extract the output of the hidden layer. Subsequently, we lower the dimensionality of this output using the *t*-SNE algorithm described above.

weakly associated with the relevant output units (i.e., the determiners). We explore these two alternatives by recording the activity of the hidden layer when presented with the words that the network encounters during training. Subsequently, we reduce the dimensionality of these representations for visualisation using the *t*-SNE algorithm. If the network can distinguish between the two groups in the hidden layer, it means that it has abstracted the relevant regions from the input and the low performance can be attributed to the weak connections. If on the other hand, the network is unable to distinguish between semantic categories in the hidden layer it points to the direction that the neural embeddings bias the network towards alternative, but partially consistent, solutions.

Figure 5.3c shows the activity of the hidden layer when probed with the training patterns from the English dataset. We see that the network discovers the neuronal ensembles that signal animacy from the input as it distributes the concepts according to the given semantic distinction in its latent space. Although we do not test for that directly, the low performance in the test set, in this case, should be due to the weak connections between the hidden layer and the output. In other words, if an animate word takes one of two determiners the network avoids accentuating the relevant regions as these might activate the wrong determiner during training. In the Chinese experiment (Fig. 5.3d), on the other hand, while we still cannot preclude the above possibility, the network appears to have problems in identifying the relevant regions from the input as animate and inanimate concepts are not linearly separable.

These two contrasting results suggest that semantic knowledge, although helpful, is not necessarily needed to achieve above chance generalisation in these tasks. The two networks achieve the same level of performance, without following the same semantic rule. In the case of English, the simulated 'learners' –unconsciously– separate their input by its semantics (or, at least, a semantic-like distinction), while in the case of Chinese they do not. Presumably, in the case of Chinese, the distributional input provided a better alternative to the model not based solely on semantic grounds. It has to be noted that this result is independent of whether the *t*-SNE algorithm clusters concepts by animacy. That is, even though the *t*-SNE algorithm can identify the difference between the two semantic categories, during training the model does not consider this dimension to be the most predictive one.

5.5 Study 6: Abstract / Concrete

5.5.1 Introduction

The second 'core' conceptual distinction that we examine is the one between *abstract* and *concrete* concepts. As in the case of animate and inanimate concepts, a significant amount of

neuropsychological data shows they are acquired and processed differently (e.g., Crutch & Warrington, 2005). Concretely, Warrington (1975) reports data from patients suffering from brain damage who exhibit selective impairment on concrete concepts, whereas their semantic system regarding abstract concepts remains relatively intact (see also Warrington & Shallice, 1984). Other studies have focused on the importance of concreteness as a psycholinguistic variable (Paivio, 1971), or its significance on vocabulary development (Brown, 1957).

Hill, Korhonen & Bentz (2014) sought to test several claims regarding the organisation and representation of abstract and concrete concepts. More specifically, they test claims by Paivio (1971) and Hopkins & Schwanenflugel (1993) that (a) abstract concepts have more but weaker connections to other concepts than concrete ones, (b) concrete concepts are organised in the mind according to similarity, whereas abstract concepts are organised according to association, and (c) concrete representations have a high degree of feature-based structure, whereas abstract representations do not. Using data from WordNet and the University of South Florida Association Norms (see §1.3.1), they find that abstract and concrete concepts differ along these three dimensions supporting claims of a differential organisation. Considering the above discussion on animacy (§5.4), Brysbaert et al. (2013) shows that the abstract and concrete labels are better thought of as the two ends of a concreteness continuum rather than categorical differences.

The above results suggest that there are representational differences between abstract and concrete concepts, but they do not necessarily imply that these differences manifest themselves in the distributional patterns of the words. This problem was also noted above in the case of animacy, however, there we were able to show that both in English and other languages, animacy can determine word order or the choice specific syntactic paraphrases. Given that Hill et al. (2014) find support for hypothesis (a) above, that abstract concepts have more but weaker connections to other concepts than concrete concepts, we expect concreteness to be reflected distributionally. The reason for expecting this is because, in the neural network setting introduced in §2.3.1, abstract words will tend to activate more nodes in their output layer than concrete ones, albeit with weaker activations. For now, we will proceed under the assumption that abstract and concrete concepts are not only organised differently, but their distributional patterns also vary. We examine this hypothesis and its implications in §5.8.2.

The tasks

We note in §1.5 that the experiments done by Paciorek & Williams (2015) focused on the acquisition of the semantic preferences of novel verbs. The introduction of this novel methodology (verbs instead of determiners) prompts us to examine whether the two tasks (those reported by Williams, 2005 and Paciorek & Williams, 2015) are equivalent, hence, comparable. There

are three possible interrelated differences we can find; (a) linguistic, (b) statistical, and (c) computational. From a linguistic point of view, the difference between the two experiments is transparent; instead of an NP, a VP is used as the critical phrase. While, however, this might seem like a little manipulation it has an interesting implication; in our survey above (§5.2) of learning the arguments of determiner phrases (i.e., the genders), we show that learners need prolonged exposure to the system to achieve native-like processing. On the other hand, learning the arguments of verbs should not be as hard as these semantically driven collocations should persist between languages. For example, the fact that the verb *eat* requires a [±EDIBLE] feature to be checked in its arguments should be independent of the language spoken.

In a related view, these differences are interesting in the context of *backwards* probabilities noted above §5.2. In the case of article → noun combinations, speakers of languages with articles should be more inclined to consider the backwards probabilities as a potential source of information. In the case of verb → argument bigrams, on the other hand, *any* speaker would benefit from considering the backwards probabilities. This dissociation makes an interesting cross-linguistic prediction; while in the case of nouns participants who speak a language which uses articles will have an advantage, this advantage will fade in speakers of languages which do not use articles. Indeed, Paciorek & Williams (2015, Experiment 3) find that the implicit learning effect reported in Experiment 1 persists in Polish, a language without articles. §6.3 looking at the transitional probabilities of article → noun combinations between languages presents a similar view.

Computationally, however, the changes introduced in Paciorek & Williams (2015) do not influence the structure of the learning model. The task of associating novel non-words with known nouns remains the same for the model in both cases. Moreover, while Paciorek & Williams (2015) use a *false memory* task during the testing phase, that is, the participants are asked to generalise to completely novel phrases, computationally the task remains the same as in both cases participants face two alternatives from which they need to choose. If we assume that participants abstract certain information from the input which is common to the stimuli in the training and testing phases, instead of remembering the stimuli they have seen, the model would not have any problems generalising to novel nouns.

Let us now look at a critical manipulation done by Paciorek & Williams (2015) in Experiments 1 and 4. Specifically, the authors explored whether semantic implicit learning effects persist even when the similarity between training and testing items is curtailed. Table 5.2 shows the differences between the two experiments. While the training lists were very similar in the two experiments, the similarity of each testing item to those lists varied between experiments (see also, Table 5.1). Paciorek & Williams (2015) find that the learning effect of drops from $\eta^2 = 0.29$ in the high similarity condition to $\eta^2 = 0.09$ in the low similarity condition.

Table 5.2 Examples of high- and low-similarity stimuli from Paciorek & Williams (2015). In each experiment participants saw four nouns with each verb (i.e., eight from each semantic category) and were asked to generalise to 32 novel verb → noun instances. The novel verbs *gouble* and *powter* are paired with abstract, whereas *conell* and *mouten* with concrete nouns.

	Phase	
	Training	Testing
Experiment 1 (High similarity)	gouble force powter status conell oxygen mouten calcium	gouble impact powter importance conell potassium mouten magnesium
Experiment 4 (Low similarity)	gouble force powter prestige conell oxygen mouten furniture	gouble surprise powter pride conell glass mouten bread

Note: During the training phase of each experiment participants saw the items embedded in English sentences but during testing they were only presented with two <verb, noun> alternatives (one grammatical and one ungrammatical).

The implication of these results is that participants do not consider abstract features such as *concreteness* to be relevant during the task, as if they were doing so, then the effect would be similar regardless of the semantic distance between the two sets. The above explanation does not preclude the hypothesis that the participants still base their decisions on abstract semantic features. In this case, the participants could still be guided by semantics, however, because of the specificity of the categories used (e.g., chemical elements) they base their decisions on a more constrained feature than concreteness. We use the WordNet representations to explore this hypothesis further.

As in the case of animacy, we expect that the neural embeddings would not only be able to model the performance of the participants in the tasks but also that they would be similarly impacted by the semantic distance manipulation.

5.5.2 Materials

Paciorek & Williams (2015, Experiments 1 & 4)

We construct two abstract-concrete datasets from the stimuli used by Paciorek & Williams (2015) using the same method as above (§5.4.2). Each dataset is split between non-overlapping

sets of training and testing stimuli. For each semantic distinction (i.e., abstract and concrete) the participants see eight items during training (four with each determiner) and are subsequently tested on 16 novel ones (eight with each determiner). There is one difference between our design and the one used in the behavioural experiments. In the original experiments, in some of the sentences the nouns were preceded by a determiner (e.g., *gouble the force*). Again, the model is not designed to account for this behaviour; however, in one experiment (Experiment 2) Paciorek & Williams (2015) found significant effects even after the removal of the determiner which indicates that the participants might not use such syntagmatic cues during the experiment. A complete description of the stimuli used can be found in §C.2.

5.5.3 Results and discussion

Figure 5.4 shows a two-dimensional projection of the stimuli used in Paciorek & Williams (2015). While the scales are meaningless in t-SNE as the algorithm chooses a random starting point, we see that in the case of the high-similarity dataset (Fig. 5.4a), the datapoints (i.e., the words) are more concentrated around the cluster centroid (i.e., the mean value of the cluster). In the case of the low-similarity dataset, we observe not only greater dispersion but also sub-groupings within the dataset. Using t-SNE instead of a variance based dimensionality reduction method, we can discover both *local* and *global* clusterings within our data. In this case, the algorithm discovers two clusters globally (abstract and concrete) but more local clusters. For example, the *cream*, *honey*, *chocolate*, *bread*, *meat*, and *wheat* cluster, which we can call *types of food*, is detached from the rest of the concrete stimuli rendering any comparisons harder for the learner. We have already argued that the t-SNE solution does not necessarily predict the performance of the participant during the testing phase as the model is tied to its initial weights and the training phase. However, the topological characteristics of the stimuli in Fig. 5.4 suggest that we should expect lower performance during the testing phase.

The generalisation gradients do not provide any useful information apart from the fact that the model quickly learns the high similarity system (peaks at 80%) while showing only a mild preference towards the grammatical alternatives (peaks at 60%) in the low similarity case. Instead, we provide the estimates of the network activations at the epoch which maximises the fit to the reported behavioural data. Figure 5.5 presents the activations of the model averaged by learner at the corresponding epochs for each model (Epoch 15 for the high similarity, Epoch 17 for the low similarity). For comparison, we also plot the behavioural results reported in Paciorek (2013) on the same dataset. While all the effects appear to be significant in the models' estimations (something that did not happen in the original human data), qualitatively, we observe that the predictions of the model are quite close to the human performance. Concretely,

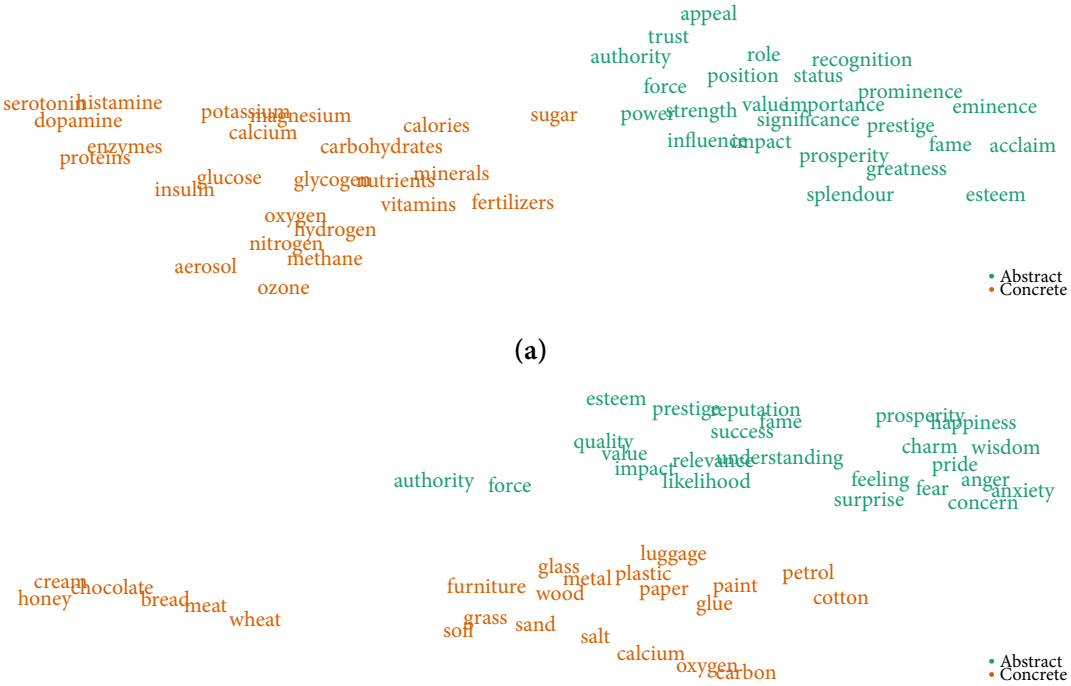
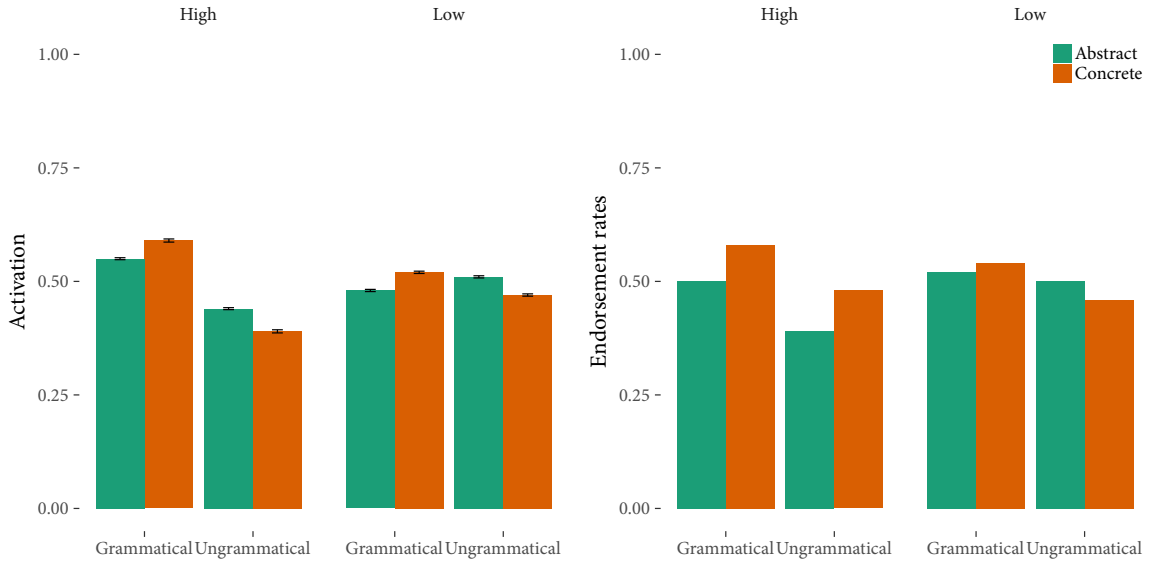


Figure 5.4 Two-dimensional projection of the stimuli used in Paciorek & Williams (2015). (a) Projection of the high-similarity dataset (Experiment 1), (b) Projection of the low-similarity dataset (Experiment 4). The high-low similarity dissociation is apparent as (a) the variance is smaller in the high-similarity case, and (b) the *t*-SNE discovers more independent subgroupings in the low similarity (e.g., bread, meat, wheat) dataset that might bias learners towards erroneous generalisations.

the model selects *concrete* words with more certainty in the high similarity dataset along with the fact that the grammatical alternatives are significantly higher than the ungrammatical ones. In the low similarity case, on the other hand, the effect vanishes for abstract concepts (as in the behavioural data), retaining somewhat a difference in the concrete concepts.

Looking at the activation of the hidden layer of the network when we present it with the stimuli from the training set sheds some light for the low similarity dataset. Figure 5.6 shows the activations for the two datasets (high and low). In Fig. 5.6a we see that the network quickly learns to implicitly distinguish between abstract and concrete concepts, enabling better generalisation. On the other hand, in Fig. 5.6b we plot the activation of the hidden layer when given the vectors of the low similarity training set. We see that even after 2000 epochs, the model cannot implicitly classify the training examples as either abstract or concrete.



(a) Model predictions **(b)** Behavioural results

Figure 5.5 Model predictions at the best epoch (High similarity: Epoch 15, Low similarity: Epoch 17) for the stimuli used in Paciorek & Williams (2015). We obtain the point estimates by maximising the fit of the predictions to the reported data (see §5.3.2). The behavioural results are reproduced from Paciorek (2013).

Presumably, the distributional features contained in the vectors are markedly different between the training and the testing sets, so the model learns some irrelevant, yet mildly predictive from its point of view, function. This result suggests that human learners, when presented with a low train-test similarity dataset, would be more prone to focus on irrelevant cues.

In §5.4.3 we noted that even though the hidden layer could not distinguish between animate and inanimate concepts, the network performs very well during training. In the high similarity case, on the other hand, the network both distinguishes internally between abstract and concrete concepts and achieves high levels of generalisation. Looking at Fig. 5.6a one more time, we first see that the network distinguishes between the two groups rather quickly. Secondly, the division between semantic groups is more clear in this case compared to Fig. 5.3c. Based on these results, we argue that the network is more certain about the regions that predict which output unit is the correct one and has no problem placing larger weights there.

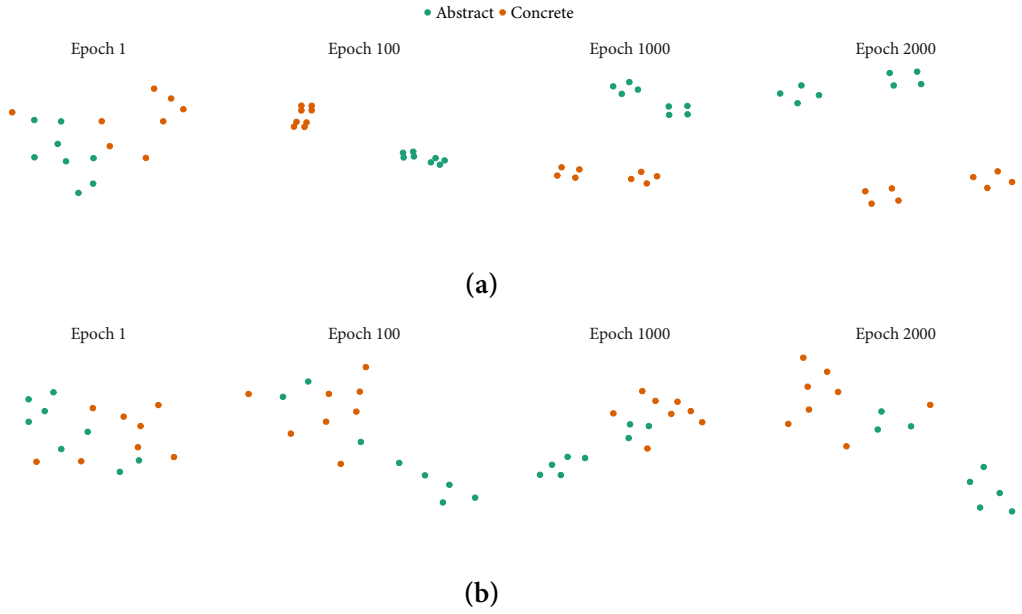


Figure 5.6 Two-dimensional projection of the activation of the hidden layer given the training stimuli used in Paciorek & Williams (2015). (a) Projection of the hidden layer given the high-similarity dataset (Experiment 1), (b) Projection of the hidden layer given the low-similarity dataset (Experiment 4). The three words in the low similarity dataset that cluster with the concrete concepts instead of the abstract were *authority*, *force*, and *value*.

5.6 Study 7: Perceptual features

5.6.1 Introduction

An alternative explanation we have not explored so far is whether the behavioural effects observed in the test sets are due to the *perceptual* similarity of the stimuli chosen than of some semantic distinction. In the two short surveys before each simulation, we have argued that the semantic implicit learning effects reviewed so far can be attributed either to some categorical dissociation (or, at least, divergent representation) of the corresponding concepts in the mind or different distributional patterns between the words of each category. Treating, however, the semantic implicit learning tasks as categorisation (see §5.2), it would not be unreasonable to consider perceptual similarity as a potential factor driving the performance of participants. Following Goswami (2008), in that *perceptual* similarity can lead to *semantic similarity*, participants might be utilising information completely independent of language to perform these tasks.

Categorising by *perceptual similarity*, and, in particular, *size similarity* has been a central idea in Gestalt psychology (Wertheimer, 1923). Under this account, in the experiment done by Williams (2005) participants might have noticed that some animals are larger or smaller than others (and similarly for the inanimate concepts) and based on that they were able to generalise to a limited extent. While a similar argument is harder to make in the case of the abstract vs. concrete distinction, recent evidence (Yao, Vasiljevic, Weick, Sereno, O'Donnell & Sereno, 2013) has found that abstract concepts are also categorised by some notion of size (e.g., *paradise* vs. *rumour*, where *paradise* is considered 'larger' than *rumour*). Unfortunately, since the McRae norms do not contain the nouns used in the experiments, it is hard to evaluate this hypothesis directly. In §5.8.2, we explore the extent to which such information differs between semantic representations based on feature norms and those generated by looking at the co-occurrence patterns of words.

Chen et al. (2011) and Leung & Williams (2012) have explored this possibility directly by including *perceptual* distinctions such that whether a concept is larger or smaller than a dog or whether the object in question when it appears on the screen is large or small. Chen et al. (2011) in one experiment performed on speakers of Mandarin Chinese report that a system based on the former distinction was unlearnable causing problems to the hypothesis that perceptual characteristics are responsible for the performance of the participants in these tasks. Furthermore, Leung & Williams (2012) report similar results from English speakers for the latter distinction (the object that appears on screen is large or small) giving further support to the null hypothesis.

The question that will concern us in the present study reduces as to whether the distributional patterns of words somehow encode perceptual characteristics such as size. According to Paivo's *dual coding* theory (Paivio, 1971) information acquired from the sensory input remains distinct from that acquired from the linguistic input. In this light, we would expect 'perceptual' information such as size *not* to be reflected in the distributional patterns of words. Indeed, recent studies exploring the interaction between distributional and experiential information suggest this to be the case. Andrews et al. (2009) show that distributional and experiential sources differ as they encode contrasting information. Subsequently, they show that a model which incorporates both distributional and experiential information performs significantly better than simpler models in a battery of cognitive tasks such as word similarity judgements and lexical substitution errors (also Andrews et al., 2014). More recently, Hill & Korhonen (2014) combined distributional representations such as those used here with perceptual information to provide a better fit to behavioural data. Perceptual information in this context comes from image descriptions coming from publicly available online datasets. Such a model provided a better match when evaluated against the Nelson norms in that it provided similar

output. These results support the idea outlined early on (§1.3.1) that the semantic system comprises many distinct sources of information. On top of that, however, the results from Chapters 3 and 4 show that different tasks or processing needs might favour one of those streams.

5.6.2 Materials

We construct two perceptual similarity datasets focusing on the size dimension from the stimuli used by Chen et al. (2011, Experiment 3). The semantic distinction in that experiment was whether the target concept was larger or smaller than a *dog*. Using the stimuli from that experiment, we construct the dataset for the Chinese simulations using the method in §5.4.2. We also use the same stimuli translated into English to simulate the effect perceptual similarity would have on English speakers. This choice has the added advantage that now the two experiments are equivalent permitting direct comparisons. A complete description of the stimuli used can be found in §C.3.

5.6.3 Results and discussion

Figure 5.7 shows the two-dimensional *t*-SNE solutions for the stimuli used in the two experiments. Despite some local clusterings (e.g., *donkey*, *horse*, and *camel* in the English case), the stimuli are rather randomly interspersed in the semantic space. As mentioned above, however, the (in-)ability of the *t*-SNE algorithm to find the intended semantic clusters does not necessarily predict the network's performance in these tasks. Figures 5.7c and 5.7d show the performance of the network on the two test sets. Evidently, in neither of the two languages was the model able to predict any learning effects despite some minuscule effects towards the end of the training phase. These results point to the direction that although 'core' conceptual distinctions such as *animacy* or *concreteness* seem to be reflected in the distributional patterns of the words, perceptual characteristics are not.

Both the effects predicted from the model and the theoretical motivation described above (see Paivio, 1971) support this latter point. A sceptic's counterargument, however, would be that the cutoff point where something would be considered as either large or small is quite arbitrary in these experiments. While we agree with this point, looking at the two-dimensional projections in Fig. 5.7, we note that it is hard to find a non-associated, non-arbitrary set of words that clusters according to size. Notice, for example, the *types of insect* cluster at the top of Fig. 5.7b. If we were to find such clusters that maximise their similarity along some perceptual dimension, we would most likely find ourselves choosing a congruent semantic distinction which might be reflected in the distributional patterns of words. This effect might

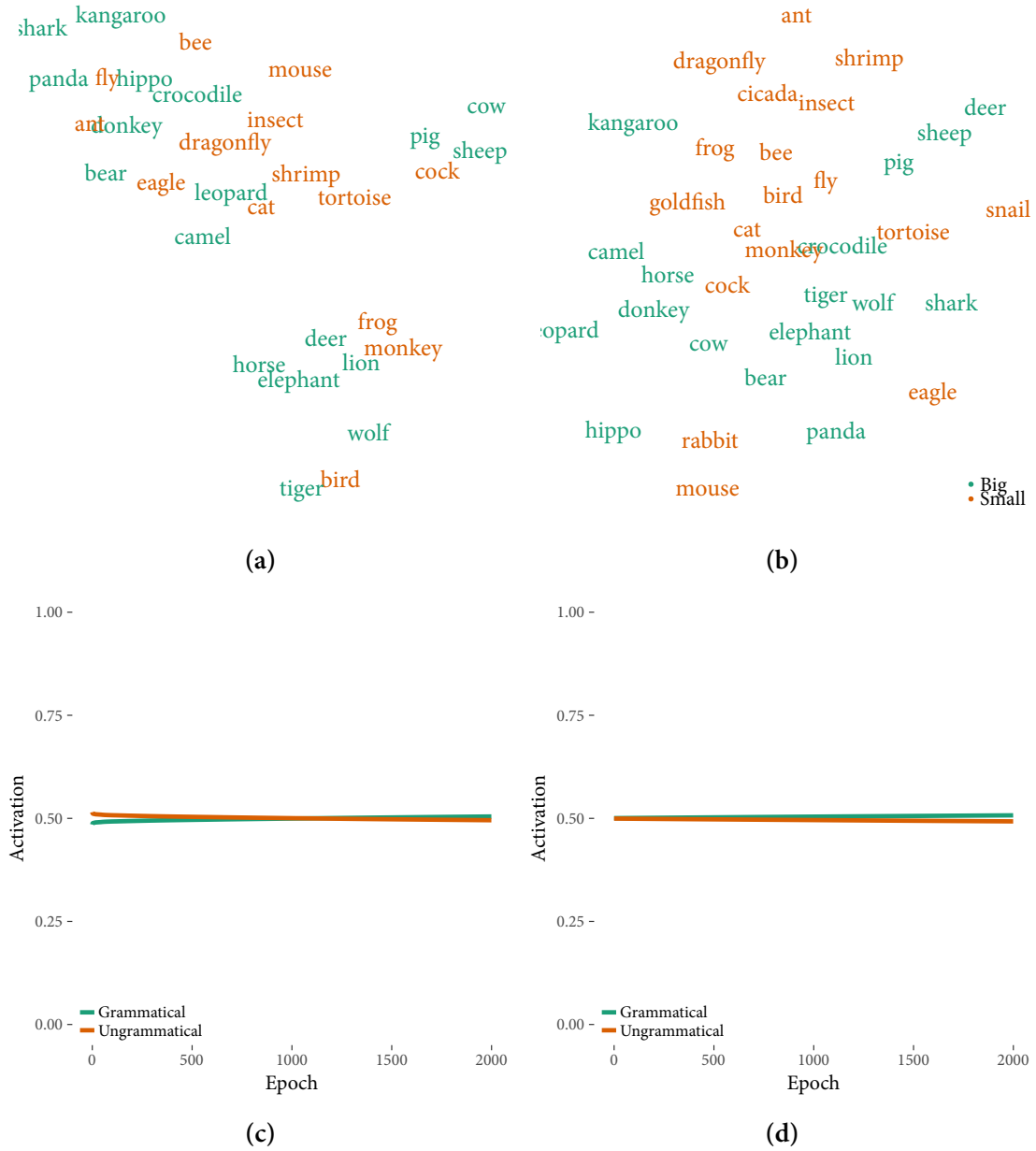


Figure 5.7 Two-dimensional projection of the stimuli used in the size experiments. (a) Projection of the words used in the English experiment, (b) Projection of the words used in the experiments with speakers of Chinese (Chen et al., 2011, Experiment 3). Since Leung & Williams (2012) did not provide their dataset, we use the English translations of the Chinese words used by Chen et al. (2011, Experiment 3). In (b), the words were translated to their English equivalents only for illustration purposes; for more details on the Chinese datasets see Appendix C.

be due to the correlational structure exhibited by the physical world (cf. Rosch, 1978) where we expect similar things to possess similar semantic features. In §5.8.2 we explore further the extent to which such perceptual characteristics might be reflected in the distributional patterns of words relating them to the priming results in §3.4.

5.7 Study 8: Language-Specific distributional cues

5.7.1 Introduction

In this last set of simulations we ask whether perceptual features of the input which are somehow manifested in linguistic usage, as in the case of the Chinese classifier system, can drive performance in SIL tasks. The results presented in the previous section suggest that the perceptual characteristics of the stimuli alone are unable to explain the generalisation patterns in the case of a relative size distinction. Furthermore, we have argued that, *generally*, the contextual distribution of each word should not contain any information regarding its perceptual characteristics. We base this argument on Paivio's dual coding theory (Paivio, 1971) which supposes different pathways for concepts or features acquired through experience and ones acquired through language. In the special case we review here, however, these two pathways are merged as words with different perceptual characteristics have distinct contextual distributions. While, in theory, DSMs should be able to capture these effects, taxonomic models, such as WordNet, would fail to predict any learning effects as they do not encode neither perceptual nor distributional information (see Table 5.1). Moreover, as we have seen above, perceptual information alone cannot explain the results of the above experiments.

Let us take a closer look at the classifier system of Chinese and examine whether DSMs would be able to capture the fact that nouns pattern differently according to their perceptual features. Both in Mandarin and Cantonese Chinese, quantifier phrases (QP) contain one particle between the quantifier and the noun called the classifier (see Examples). These classifiers resemble articles in Romance and Germanic languages in that they collocate with different nouns. Unlike Romance and Germanic languages, however, this collocation is mostly determined by the semantics of the quantified noun (Jiang, 2017). Several researchers (Lyons, 1977; Mithun, 1986; Pulman, 1978) have argued that classifier categories are grounded in perceptual properties of the input, and, more specifically they can be distinguished by *shape*, *size* or some arbitrary *semantic feature* (e.g., 扇 'a leaf-shape thing') (Gao & Malt, 2009). Apart from their co-occurrence patterns, their number is much higher than the average number of articles in a Romance or Germanic language (Chao, 1968 estimates that there are about 50

classifiers in Chinese, whereas Zhang, 2007 puts this number at over 900) which, subsequently, lowers the number of possible classifier → noun combinations.

(5.15) 三 只 貓
three CL cat
'three cats'

(5.16) 這 只 貓
DEM CL cat
'this cat'

(5.17) 這 三 只 貓
DEM three CL cat
'these three cats'

Even though Chinese nouns pattern differently according to their semantics and, in particular, semantic properties, the question as to whether Chinese classifiers bias language processing has been a contentious issue (Jiang, 2017). Only a handful of studies (Saalbach & Imai, 2007; Schmitt & Zhang, 1998; Srinivasan, 2010; Zhang & Schmitt, 1998) have tested directly whether Chinese and non-Chinese speakers would process differently similar input. In one instance, Srinivasan (2010) has looked at whether knowledge of classifiers affects performance in a non-semantic task. In a visual search task, participants were instructed to count how many times a specific object occurs in a picture. Crucially, in the same picture apart from the critical object there were several distractors which either take the same or different classifier as the target in Chinese. Srinivasan (2010) found that speakers of Chinese showed greater interference in the trials were the distractors took the same classifier as the critical nouns compared to the unrelated condition. Crucially, their times were higher than those of speakers of Russian and English showing a language-specific bias.

On the other hand, Saalbach & Imai (2007) show than on a battery of tasks involving German and Chinese speakers responding to taxonomic and classifier relationships, Chinese speakers did not behave any differently than German ones apart from some minor advantages (see also Imai, Schalk, Saalbach & Okada, 2014). Saalbach & Imai (2007, exp. 4) performed a speeded picture matching task in which participants see a cue word and then they have to decide whether a subsequent target picture matches that word. Critically, when the target picture was associated to the cue by a classifier relation (i.e., both the cue word and the word from the target picture take the same classifier) the Chinese speakers showed no advantage. On the basis of this, the authors argue that in priming tasks classifier relations do not influence reaction times. We come back to this point in the discussion showing that the dataset used by Saalbach & Imai (2007) might have biased participants. In sum, however, the above

results suggest that at best we should expect only a minor effect of classifiers on semantic representations as captured by DSMs.

5.7.2 Materials

We construct both the English and the Chinese datasets using the stimuli from Leung & Williams (2014). Leung & Williams (2014) performed a speeded reaction time task (more details in §1.5) on speakers of Cantonese Chinese and British English. However, the distribution of Cantonese classifiers differs from that of Mandarin Chinese which is the 'dialect' of the Chinese corpus we trained in §4.5. This difference can become problematic in the present case, as the semantic representations generated above assume the distributional properties of Mandarin Chinese. If the distribution of determiners differs from that of Cantonese, this can skew our results as we would only rely on the perceptual characteristics of the input.

We mitigate this problem by using the MDBG dictionary[11] and selecting those words that in *Mandarin* Chinese take the same classifier for long and flat, discarding the rest. Using this procedure, we end up with a dataset of 68 training words (17 for each determiner) and 20 items in the test phase (five for each classifier). As in Williams (2005) there is an overlap between the train and test sets as the participants have already seen the critical nouns with another determiner during training. Regarding the English dataset, we adjust it accordingly, retaining only the words which are equivalent to those used in the Chinese experiments. Details on the actual stimuli both in English and Chinese used during the experiments can be found in Appendix C.

5.7.3 Results and discussion

As in the previous experiments, we start by looking at the two-dimensional projections of the words in the English and Chinese distributional spaces respectively. Figure 5.8a plots the English distributional vectors of the words used in the experiment. Interestingly, the words pattern quite differently in English (also consider the arbitrariness of the semantic distinction), which prompts us to consider whether the system would be learnable even in English. However, looking at the smaller clusters identified by the *t*-SNE algorithm, we see that the distributional vectors possibly contain other sources of information not directly related to the thin–flat distinction. For example, *celery*, *sausage*, *carrot*, and *banana* form one group, or *poster*, *photo*, *postcard*, and *painting* another. Distributionally, these words should be similar, but irrelevant to the semantic distinction at hand. Spurious relations, such as *bedsheet* and

[11] https://www.mdbg.net/chinese/dictionary

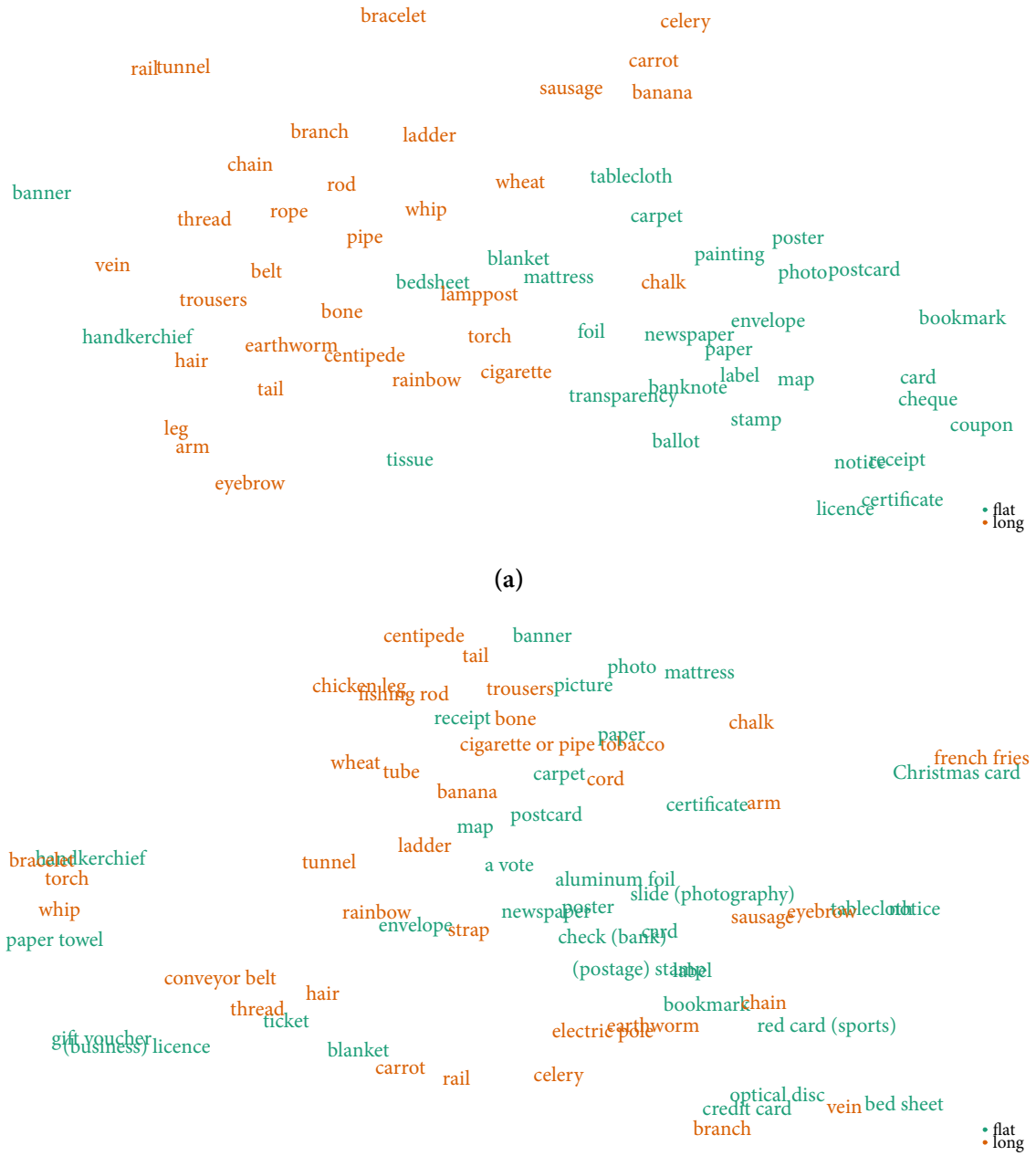
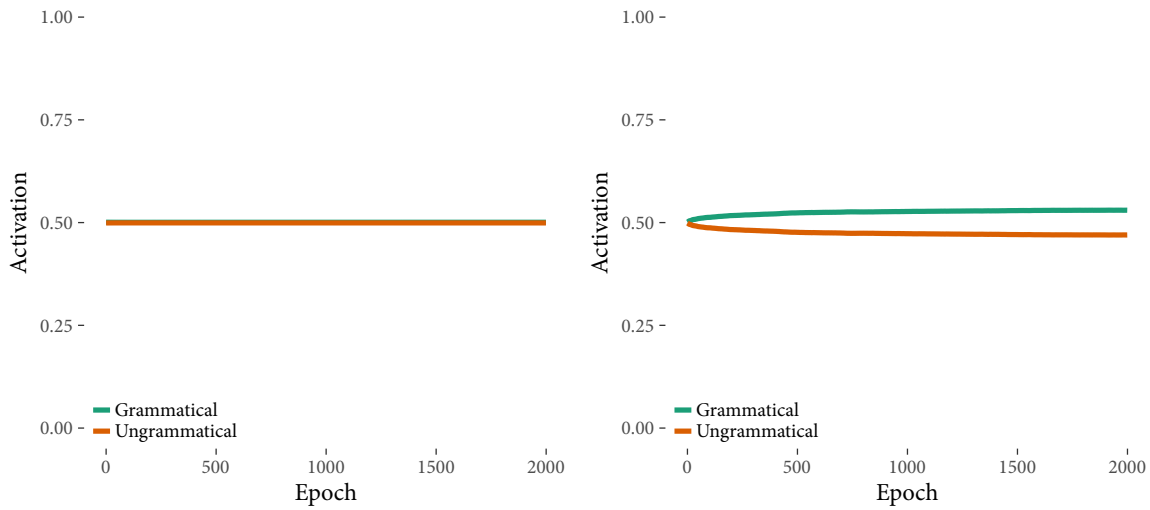
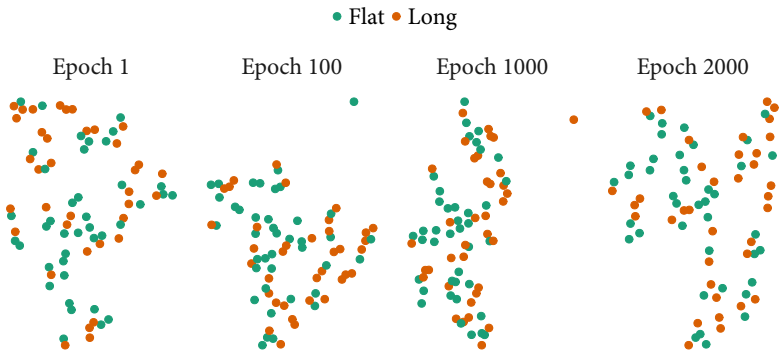


Figure 5.8 Two-dimensional projection of the stimuli used in the long-flat experiments. (a) Projection of the English words, (b) Projection of the Chinese words. The Chinese words are translated to their English equivalents only for illustration purposes; for more details on the Chinese datasets see Appendix C.

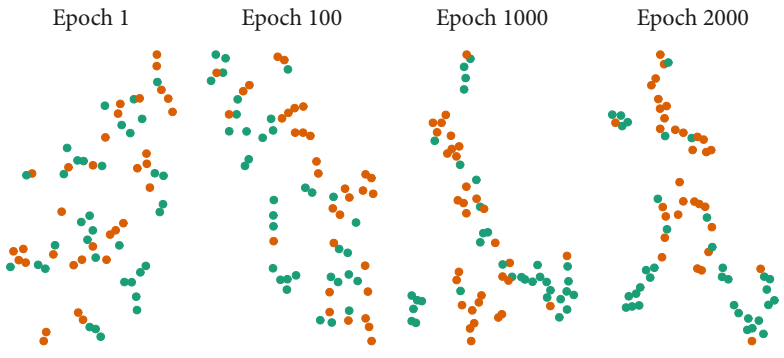


(a) Generalisation performance (English)

(b) Generalisation performance (Chinese)



(c) Activation of the hidden layer (English)



(d) Activation of the hidden layer (Chinese)

Figure 5.9 Generalisation gradients for the two long–flat experiments. Figures 5.9a and 5.9b plot the generalisation performance of the model for English and Chinese, respectively. Furthermore, Figs. 5.9c and 5.9d plot the activation of the hidden layers given the words in the training phase.

lamppost can be explained by the fact that compressing the words in two dimensions will result in information loss.

The Chinese vectors, on the other hand, do not pattern that well. While some by-classifier clusterings can be seen, it might seem that the system would not be easily learnable by a speaker of Chinese. For the moment, we will refrain from considering alternative explanations for this behaviour (e.g., the quality of the corpus, or our lemma selection procedure) as the network might still be able to recover the intended distinction. The reason for believing that during training a different picture might emerge is because even though there is more of a clear distinction in English than in Chinese, in English the words pattern in smaller clusters. This configuration points to the direction that the vectors might lack the sort of information that would enable accurate generalisation (e.g., flatness). For example, during the behavioural experiment, we might conjecture that the learner entertains multiple hypotheses at the same time, which do not necessarily guarantee high performance during testing. On the other hand, if the Chinese vectors, despite their dissimilarities, contain a common semantic feature by appearing in similar contexts, then generalisation would be easier.

Figure 5.9a plots the generalisation gradients of the English learners in the test set. Interestingly, the model is unable to generalise to the novel determiner → noun collocations despite 'knowing' that the nouns pattern differently. Figure 5.9c corroborates these results, as the model even after 2000 epochs is unable to find a pattern that reliably predicts the correct determiners. Figure 5.9b presents a more encouraging view for the Chinese results. Despite the model levelling at about 55% its performance differs from chance. This result is in line with those reported by Srinivasan (2010) in that 'categorisation' using Chinese classifiers can only exhibit minor, yet consistent, effects. The activations of the hidden layer given the training set complement the above hypothesis. The Chinese distributional vectors, even though they did not seem to pattern by the classifier they take in the *t*-SNE solutions, they seem to contain that sort of information as the network distributes them differently to make its predictions. Even though this classifier distinction is more 'noisy' when compared to the *animacy* or *concreteness*, it manifests as a reliable solution for the model.

The above results are intriguing considering the inability of Saalbach & Imai (2007) to obtain significant differences between speakers of German and Chinese in semantic tasks involving a classifier distinction. Taking a closer look at the dataset of Saalbach & Imai (2007, Table 1, p. 490) (also in Table C.13), we see two –possibly interacting– explanations; firstly, the stimuli chosen in that dataset are too dissimilar. Considering the above results, as well as the results obtained by Srinivasan (2010), we expect Chinese classifiers to marginally bias the semantic space. This bias has the effect that some dimensions of these words are closer together in the high-dimensional space than expected, causing the learners to show some preference

for same classifier nouns. Using the representations from our DSM, we examine whether the similarity of the words used in the behavioural experiments differs in each category (classifier, taxonomic, and so on). In our model, as above (pg. 77), we measure the cosine distance between the cues and either the controls or same-classifier words. We find no significant differences in the similarities between the cues and either category. On the other hand, the effect is boosted in the taxonomic and thematic relations, which are significantly more similar ($F(3, 52) = 22.533$, $p < 0.001$). Post-hoc analyses using Tukey's test revealed that there was no difference between classifier and controls or between taxonomic and thematic relations but the differences between the other relations were significant as suggested by the behavioural data. A related second reason is that Saalbach & Imai (2007) use a limited number of stimuli for each classifier set (one, two, or three items). In light of the above results, even if the distributional representations somehow reflect the classifier distinction, the learning model will have very few instances from which it can generalise.

5.8 Discussion for Studies 5–8

The results of the above simulations suggest that only representations based on the distributional patterns of words can model tasks of semantic implicit learning. We arrive at this conclusion by looking at various datasets using different semantic manipulations and speakers of different languages. Concretely, we start by looking at behavioural experiments using 'core' semantic distinctions such as *animacy* and *concreteness*. Both the taxonomic and a distributional semantics model can predict the learning effects there, although the simulations show that the DSM does so in a more human-like way. Subsequently, we look at datasets that explore the learnability of systems based on perceptual features. Interestingly, this manipulation failed to yield learning effects both in humans, in the behavioural experiments, and in the computational simulations. Finally, we look at the results of a system based on a co-occurrence rule in Chinese. In these experiments, speakers of Cantonese Chinese showed greater learning effects than their English counterparts, presumably because the distributional rule was found in their L1. Since taxonomic models assume that speakers of different languages share their semantic space, only a DSM trained on raw linguistic data can pick up that rule. These results are significant as they implicitly propose that semantic implicit learning does not reflect semantic organisation but some distributional biases of the human participants.

In what follows we discuss various topics that we have highlighted throughout this chapter. We structure our discussion around three axes. Firstly, how can the results of the above simulations inform a computational theory of semantic implicit learning? Secondly, what

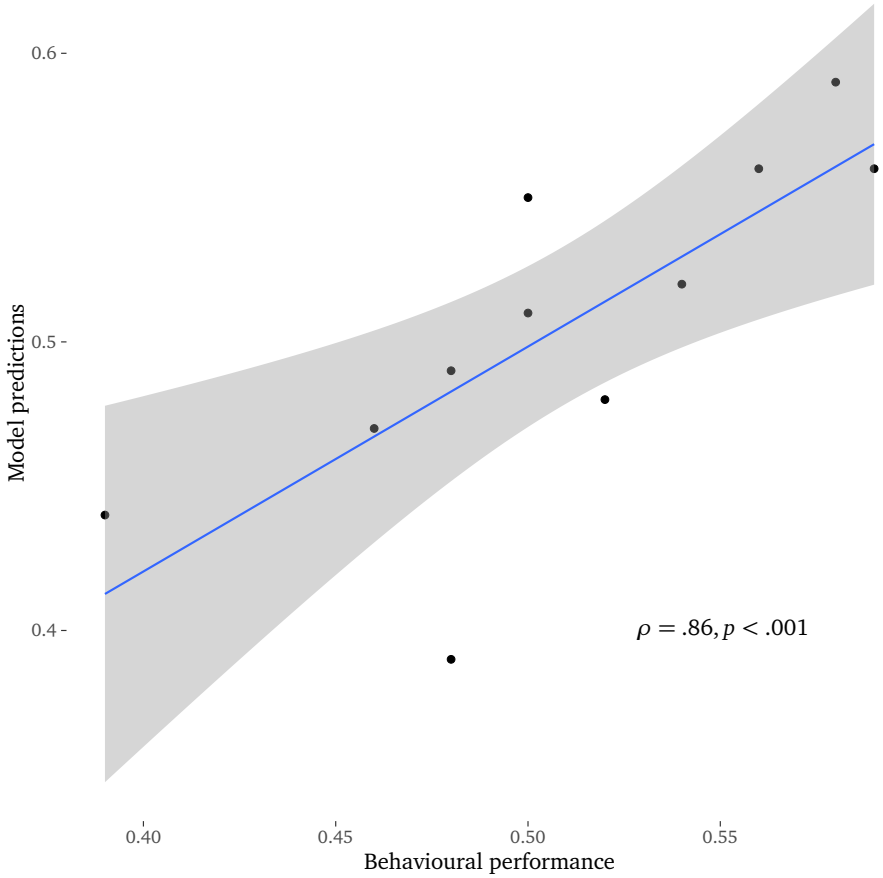


Figure 5.10 Correlation between observed behavioural performance on the 2AFC tasks and corresponding model predictions. We gather the behavioural estimates from the experiments reviewed above (except for Leung & Williams, 2012, and Leung & Williams, 2014, which do not use a 2AFC task) and extract the model predictions using the point estimation method outlined in §5.3.2.

kind of semantic features can we expect to be seen in large linguistic corpora? Thirdly, we discuss various computational issues that we identified throughout the simulations.

5.8.1 A theory of Semantic Implicit Learning

The results obtained in this chapter point to the direction that semantic implicit learning need not be considered semantic at all. In fact, the patterns obtained in the behavioural datasets can be recovered through the distributional characteristics of the stimuli used in the experiments instead of abstract semantic features. Indeed, looking at the 2AFC tasks alone (i.e., excluding Leung & Williams, 2014, which uses a speeded reaction time task), we achieve

a strong correlation ($\rho = .86, p < .001$) between the model estimates and the behavioural performance (Fig. 5.10). An interesting remark we can make about this correlation is that we achieve a good fit to the human performance *without* directly fitting our models to the human results. In other words, the models are not *trained* to match the human performance in the SIL tasks, but to predict semantic priming. We only fit their estimates post experimentally, but even in that case, the generalisation gradients do not show a lot of variance (they usually level after a few dozen epochs).

Linguistic relativity?

Insofar as we consider the semantic representations constructed by the DSMs as a proxy of the true semantic representations, it might follow that speakers of different languages conceptualise the world around them slightly differently. However, classic work in cognitive psychology has determined that despite surface linguistic differences, humans conceptualise the world in a very similar way. As we have argued above (§1.5), in the domain of colour perception, Berlin & Kay (1969) using data from 20 different languages identified a 'universal' evolutionary pattern in colour naming. For example, *all* languages contain terms for black and white; subsequently, the order by which colour have a specific term in the language is *red*, *green*, *yellow* and so on. Similarly, Eleanor Rosch (Heider, 1972; Heider & Olivier, 1972) found that even though the Dani people in Papua New Guinea lacked the terms for any colour other than *dark* and *light* (cf. black and white), they were able to categorise objects by colour for which they had no word. The above results led Rosch to assert that it is not the structure of each language that determines conceptual organisation but a pressure for efficiency, on the one hand, and common world knowledge, on the other (but see some recent results from Cibelli, Xu, Austerweil, Griffiths & Regier, 2016).

Under this light, our proposal that specific distributional patterns might give rise to different conceptual structure between speakers of different languages might seem problematic. Indeed, in the limit, this proposal predicts that speakers of different languages will have distinct semantic spaces and process incoming input differently. While this topic is quite contentious in cognitive psychology (Brody, Gumperz & Levinson, 1998) and any proposal would be met with severe criticism from the other side, we find two middle-ground solutions to this problem. Firstly, as we noted above, despite the fact that overall speakers of different languages share a semantic space, minor effects attributed to distributional knowledge can be found. Secondly, in §7.2, we outline a study in which we transform distributional representations to semantic representations containing information about semantic relations. Under this account, the distributional knowledge might simply provide a different starting point for

speakers of different languages which through exposure to the physical world is refined ending up in a similar semantic space.

A role for phonology

As noted at the beginning of this chapter, our method does not take into account the phonological representations of the words in question. This happens because we encode each word as a one-hot vector (i.e., a localist representation), making it orthogonal to every other word. While we argued that this should not be an issue in the present context where we look only for the contribution of distributional semantics to the implicit learning tasks, it remains an open question whether a model which looked only at phonological information would be able to explain the results. One can imagine that upon encountering 'gi dog', 'gi drill', 'gi dark' a participant might use such cues to limit their search space and arrive at an ad-hoc categorisation such as 'things that start with the letter d' (Tenenbaum, 1999, for an example in number learning). From a modelling perspective there are various ways we can deal with this issue; firstly, following connectionist models on word-reading (Seidenberg & McClelland, 1989) one can imagine that the hidden layer receives its input from two distinct streams (one semantic and one phonological) and its output activation pattern depends on a (non-)linear combination of the two. Secondly, another possibility would be to use phonologically motivated input vectors where each unit represents a particular phonological unit (McClelland & Elman, 1986) or that they are sampled from different distributions (Jones & Mewhort, 2007). In the behavioural domain, there have been recent contributions for the interaction between semantic and phonological knowledge (Ouyang, Boroditsky & Frank, 2016). While these are reasonable objections and deserve a more rigorous examination, in the present chapter, we are only interested in the contribution of semantic knowledge in such tasks, leaving an exploration of phonological effects for §6.2.

Is this the whole picture?

A related question concerns the limits of the learnability of such systems. Could this mean that any semantic distinction can be learnable? Potentially yes, but with considerable constraints; firstly, the introductory discussions for every study assert that the semantic distinction has to be somehow reflected in linguistic usage. This property quickly constrains the space of possibly learnable systems as there is a limited number of ways words can be combined in a language to yield different distributional patterns. However, the results given here are consistent with the idea that novel semantic distinctions can be formed. Since the network constantly learns from its input, if we transform the input in such a way to highlight novel semantic distinction

these should, in theory, be learnable. Recall at this point the discussion in §3.1, where Barsalou (1983) recognises the potential that through usage, *ad hoc* categories (e.g., things to take in the event of fire) should constitute a more natural category for a particular speaker. Our results support this notion, as repeated sentences in a corpus that contain the objects that one takes in the case of fire will bring the corresponding concepts closer in the semantic space, resulting in faster processing and learning.

Secondly, the results from the English speakers in §5.7 show that the semantic distinction to be learnt has to preclude more specific, and more intuitive hypotheses as these might bias the learner away from the 'true' distinction. Tenenbaum & Griffiths (2001) explain this behaviour in terms of *Bayesian* priors. In short, the learner *will* prefer the hypothesis that (a) fits the data best, and (b) is more intuitively probable. While how we define intuition (or prior probabilities in the Bayesian context) is a contentious issue in cognitive science (e.g., the discussion between Bowers & Davis, 2012 and Griffiths et al., 2012), for the English concepts being long might be a less probable categorisation than being an insect, even though they might fit a portion of the data the in the same way. Considering the correlational structure of concepts in the world (Rosch, 1978), it might be hard to construct such an artificial category, where the subclusters are not more probable.

Do all the 'learners' learn in the same way?

The final question we are going to examine in this section is whether all the simulated participants follow the same learning path during each experiment. Here we do not examine how individual differences stemming from L1 biases might affect participants' framing of the computational problem. In the applied linguistics literature, this sort of individual difference concerns whether participants notice the relevant variables (animacy and the determiner) or not.[12] In the present case, we explore differences in performance that stem from different initialisation of network weights, the randomised order of presentation, and the *dropout* procedure outlined in §5.3.1. All these factors can be considered as noise during the experiment. Human learners might perceive the input noisily; they might forget or misremember a particular example. We, therefore, ask whether the learners would converge to the same output despite these differences or random factors determine their performance.

Figure 5.11 plots the error and generalisation gradients of five simulated learners from the Paciorek & Williams (2015, High similarity) dataset. Interestingly, the results show that not all participants achieve the same level of performance in the experiments. Figure 5.11a plots the summed cross-entropy error for each participant by semantic distinction. Looking

[12]An important side note here would be that even if participants notice the relevant variables the learning can still be considered implicit as they do not notice the relation between them.

at the gradients, we see that some participants find it easier to learn either concrete or abstract concepts, whereas some others might not learn at all. These error patterns seem to also extend to the generalisation rates in the testing phase. Figure 5.11b plots the unnormalised activation of the grammatical alternatives for abstract and concrete concepts. This data shows that the participants can generalise to the extent that they managed to 'learn' the training set. Learner #4, for example, did not learn anything during training and performed completely at chance during testing (recall that these are the unnormalised activations). Participant #1, on the other hand, generalises better to concrete concepts which they managed to learn better during training.

These results highlight the fact that even in the case where participants frame the task in an optimal way (i.e., retrodiction), achieving good levels of generalisation is still quite hard. Many issues can appear such as errors during retrieval or the effect of randomisation that might prohibit participants from achieving high performance. In §7.3 we discuss the latter reason to some extent in the light of *curriculum learning*. In short, *curriculum learning* (Tsvetkov, Faruqui, Ling, MacWhinney & Dyer, 2016; also see Elman, 1993, for an early connectionist overview) assumes that there is an optimal path during training that aids learning. The idea is that if there is a target rule to be learnt, receiving random examples might prohibit the model/learner from abstracting the relevant information. If on the other hand, the training stimuli are administered in such a way that enables abstraction of the relevant information then the learner can achieve higher levels of generalisation. Our results are consistent with this view, and can potentially be explored further in future research to optimise the learners' input.

5.8.2 Linguistic usage and semantic features

Can function words influence semantics?

The simulations presented in §5.7 build on the idea that in Chinese, classifiers influence the contexts in which nouns can appear. Therefore, *thin* or *flat* objects will be clustered together more in Chinese because they appear in similar contexts modulated by the presence of the classifiers. This idea, however, is not without problems; from a computational point of view, the distribution of Chinese classifiers is similar to that of articles in many gendered languages (as in Italian) with the main difference being, as noted above, that the grammatical constructions which need a classifier are more constrained. The problem arises in languages such as Italian or French (or other languages that have a grammatical gender) where words should cluster together distributionally by appearing in similar article-contexts but do not seem to do so.

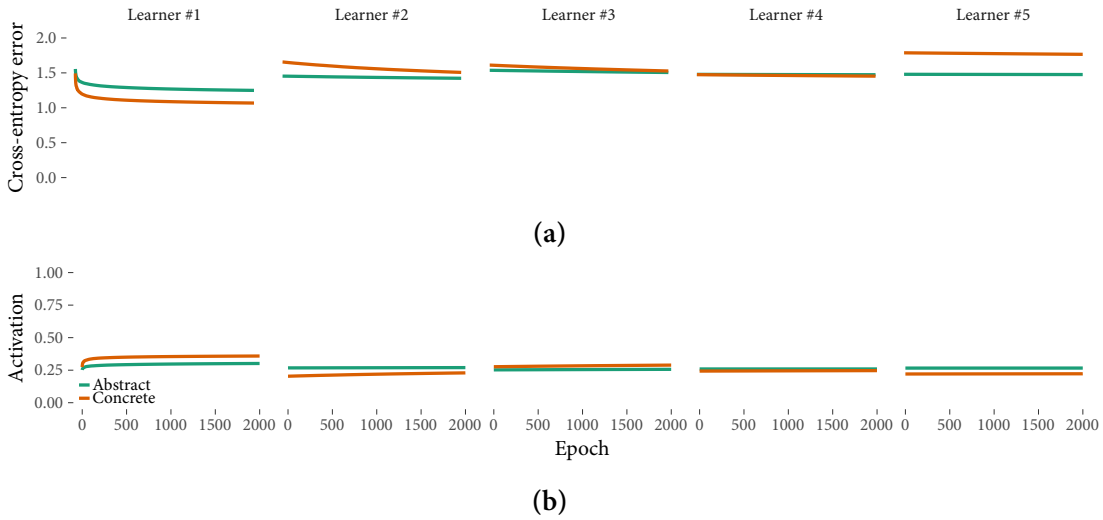


Figure 5.11 Results from five simulated learners using the Paciorek & Williams (2015) dataset. (a) By epoch error rates colour-coded for semantic distinction, and (b) the associated by epoch unnormalised generalisation rates (i.e., performance during testing).

Instead, gender assignment in such languages is predominantly based on either morphological or phonological factors (Corbett, 1991) (see also Chapter 6).

Figure 5.12 presents a two-dimensional projection of 300 Italian nouns† (100 for each article) colour-coded by the article they take. If consistent appearance with a specific article would guarantee semantic clustering, then we would expect nouns to cluster according to the articles they take instead of randomly. To put this figure into a quantitative context, we also ran a *k*-means clustering algorithm (e.g., Arthur & Vassilvitskii, 2006) on the raw distributional vectors to see if the gender subclasses could be discovered solely from contextual information. *k*-means is an unsupervised algorithm which tries to find *k* cluster centroids such that the within-cluster *inertia* is minimised. That is, given a configuration of points in space and some clusters to look for, *k*-means finds the clusters such that the within-cluster distances are minimised. Since *k*-means needs to know the number of clusters beforehand, we specified them to be 3 (i.e., the number of different articles). We can then take the cluster predictions of the algorithm and compare them to the correct gender classification. We use the Adjusted Rand Index (ARI) (Rand, 1971) to evaluate the performance of the clustering algorithm. The ARI measures the similarity between the true clustering and our predicted clustering adjusting for chance assignments. Concretely, the ARI measures the number of agreements between the true clustering and our prediction, dividing by the possible number of clusterings yielding a score between -1 and 1 (where 0 is chance). Since *k*-means is not robust to *local minima*,

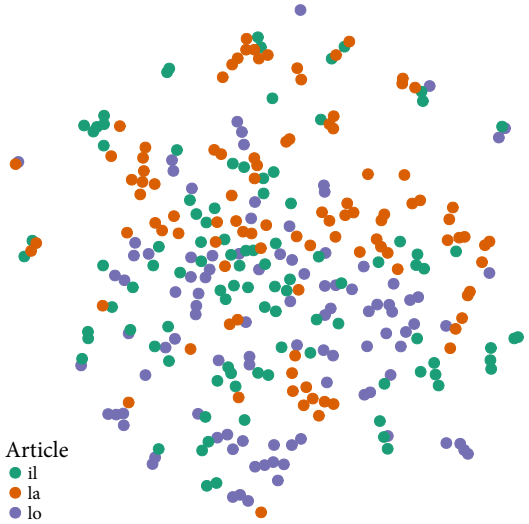


Figure 5.12 Two-dimensional projection of 300 italian nouns colour-coded by the article they take (note that *il* and *lo* are both masculine). Despite some local clusterings, which are to be expected, a k-means algorithm has failed to provide clusters which agree with the grammatical gender (see text).

we repeat the clustering 100 times and report the average. The mean ARI measure for our clustering was 0.03 which would be very close to the chance label assignments.

The results from Italian render those from the Chinese nouns all the more surprising as the Italian nouns do not seem to encode any information about the article they take. Before attempting to explain why the model behaves in such a way in Chinese, we check whether Chinese nouns do indeed cluster by classifier and do not see an artefact of the simulation. To test this, we select words from a Chinese dictionary (the MDBG used above) along with the classifier they take. By looking at their definitions, we filter out the words which can be defined by one word which also appears in the English model (e.g., the definition for 貓 is *cat* which can be found in the English model). Furthermore, we remove the classifiers with ten words or less leaving a total of 95 words (from 1805). We then train two independent k-means models (one for English, one for Chinese) on those 95 word vectors where $k = 4$ (the number of classifiers remaining). After 100 simulations the mean ARI for English was 0.08 while for Chinese 0.24 (both significantly higher than zero as revealed by one-sample t-tests $t(99) = 9.16$, $p < 0.001$ and $t(99) = 29.5$, $p < 0.001$, respectively, and significantly different from each other $t(195.94) = 12.95$, $p < 0.001$, two tailed).

Table 5.3 The Chinese classifiers used in the clustering simulations along with the number of words using that classifier. Although all Chinese nouns are associated with a classifier we end up using only 95 words because of (a) our matching procedure to English nouns and (b) our constraint to have at least ten words for each classifier.

Classifier	Mandarin Pinyin	Applies to	# elements
張	zhāng	Flat objects	13
條	tiáo	Long thin objects	17
隻	zhī	Birds and certain animals	16
個	gè	People or objects in general	66

The above results show that Chinese nouns are affected more by the presence of the classifiers than in either English or Italian. Even if that is the case, and articles or article-like particles are allowed to influence the semantic representation of a word, we would expect them to behave similarly in either Italian or German, something that we saw above does not happen. An explanation for this might be provided by Table 2.3 where we discuss various *vector normalisation procedures*. We noted there that because of how words are distributed in languages around the world (cf. Zipf, 1949), we would expect that all the words should have 'peaks' in similar points in the high-dimensional space (for example most words should have a high count for 'the'). To counter this effect, avoiding spurious correlations, we argue that we should normalise the word vectors removing the effect of high-frequency words. Take the *tf-idf* method, for example. Applied to a $\mathcal{V} \times \mathcal{V}$ matrix where every row is a word, and every column is the context, the cell M_{ij} counts how many times the word j has appeared as a context to word i. If the column j corresponds to the word *the* which can appear as context to many words this value will be downscaled by the overall frequency of *the* nullifying its effect, reflecting that it is not a very informative word.

On the other hand, the appearance of Chinese classifiers is not as consistent as articles in English or Italian. Firstly, the number of classifiers is much larger than that of articles in the examined languages. As an example, Chao (1968) estimates that there are about 50 classifiers in Chinese, whereas Zhang (2007) puts this number at over 900. Since most nouns are associated with one classifier, a corollary of the above is that every classifier is associated on average with fewer nouns than every article in a Germanic language. Thirdly, since the basic use of these words is in quantifier phrases, the number of contexts in which they appear is much smaller than that of the articles. Considering all the above, we suggest that the distribution of Chinese classifiers, although similar to that of determiners, is not the same, at least as far as DSMs are concerned.

Is concreteness reflected in corpora?

It remains an open question whether the distinction between abstract and concrete concepts is manifested in large linguistic corpora. The results from Hill et al. (2014) discussed above, on the one hand, show that concrete concepts are organised differently than abstract in the mind. In the case of *animacy*, looking at linguistic examples, we argue (§5.4) that we should be able to identify *animate* nouns based on their co-occurrence patterns. Both the results of the simulations in §5.4 and the two-dimensional projections offered therein show this to be the case in English and Chinese. However, in the case of *concreteness* different organisation does not necessarily imply different linguistic manifestation.

If we argue solely from the results of Hill et al. (2014) that this distinction would show in language usage, then we implicitly subscribe to a strong distributional hypothesis. While we have tacitly avoided this issue, there are two variations of the distributional hypothesis introduced in §2.2. A *weak* distributional hypothesis (Lenci, 2008) posits that the elements of the meaning of each word (however we define meaning) are recoverable by its distributional patterns. Under the assumption that there is a link between language distributions and semantic content we can get a better idea of the meaning of each word. On the other hand, a *strong* distributional hypothesis assumes a causal link between linguistic usage and semantic content. In other words, because the linguistic input is structured in a certain way, this drives language learners to structure their semantic space similarly. This assumption is behind the models introduced in Chapter 2 and the study outlined in §7.2.

In the present discussion, we provide further evidence from corpus analyses by finding direct links between a corpus estimate of concreteness and human judgements. To do that, we extract concreteness estimates for 8000 nouns contained in the British National Corpus and an online database as described in Brysbaert et al. (2013). Brysbaert et al. (2013) have conducted an online study gathering concreteness estimates for 37058 English words obtained from over 4000 participants. Following Rosa, Catricalà, Vigliocco & Cappa (2010), each participant had to indicate how the meaning of 300 words is acquired, by scoring each word on a five-point scale ranging from purely experience-based (concrete words) to language-based meaning acquisition. If concreteness is reflected in large linguistic corpora, then it must be something in the embeddings of concrete concepts which differs from the abstract ones.

We estimate the concreteness of words in the corpus based on the *Context Availability Model* (Hopkins & Schwanenflugel, 1993), which relies on the hypothesis that abstract words have more but weaker connections to other (concepts) than concrete words. Consider now how the neural network presented in Fig. 2.2 would solve this problem. If the average abstract word causes the transformation of the embedding layer by the context matrix to activate more units in the output but vice-versa for the concrete words, then to learn the embeddings, the

network has to learn the weights that capture this variability. We measure this variability by performing Principal Components Analysis (PCA) on the embedding matrix of the 8000 words used in the Brysbaert et al. (2013) experiment and use the *first* principal component (PC1) as our dependent variable.

Fig. 5.13a shows a two-dimensional projection of the distributional word vectors corresponding to the nouns used in Brysbaert et al. (2013). For the distributional vectors, we use the word embeddings identified in §3.4. As described above, the scale and the sign of the principal components in PCA is meaningless and is used here to explore whether there is a 'cut-off' point between abstract and concrete nouns. While there doesn't seem to be a natural line dividing the nouns, a clear pattern emerges if we colour each noun according to its concreteness estimate. Words which have been described as 'very concrete' are coloured in red while words in 'blue' are the ones described as 'very abstract'. While qualitatively one can see that there is a clear demarcation between abstract and concrete nouns, what is even more interesting is that the colour gradiant transitions from deep red to deep blue in the y-axis. This result points to the direction that the neural embeddings are not only able to capture concreteness and broad terms (say that 'mouse' is qualitatively different from 'stink') but in a human-like way.

We quantify this qualitative relationship by correlating the first principal component with the concreteness estimate offered by the behavioural study (`Conc.M` in the original dataset). Fig. 5.13b plots the correlation between the 8000 nouns found in the Brysbaert et al. (2013) dataset and the first principal component obtained from PCA. There is a highly significant monotonic relationship between the two variables (as described above the sign in PCA is meaningless) $\rho = -.62, p = 0$. Unlike Hill et al. (2014) who found a correlation between concreteness measures and word frequency, we did not find any correlation between `Conc.M` and either SUBTLEX frequency (which is included in the dataset) or with word frequencies extracted from the BNC. In any case, we included both the first principal component (PC1) and BNC per million word frequency in a linear model as predictors and concreteness as a dependent variable. PC1 remains a significant predictor of concreteness even when we include frequency as another predictor $\beta = -1.007, p = 0, R^2 = .382$.

The above discussion highlights the fact that the distributional properties of nouns in corpora encode something akin to concreteness. We find this using the *Context Availability Model* (Hopkins & Schwanenflugel, 1993) that posits that abstract words are related to other concepts in a different way than concrete ones. However, it is important to note that there is no one-to-one mapping from a single dimension in the word embeddings and a concreteness feature. Both the PCA visualisation and the strong correlation between the first principal component and the concreteness ratings warrant this conclusion. The first principal component

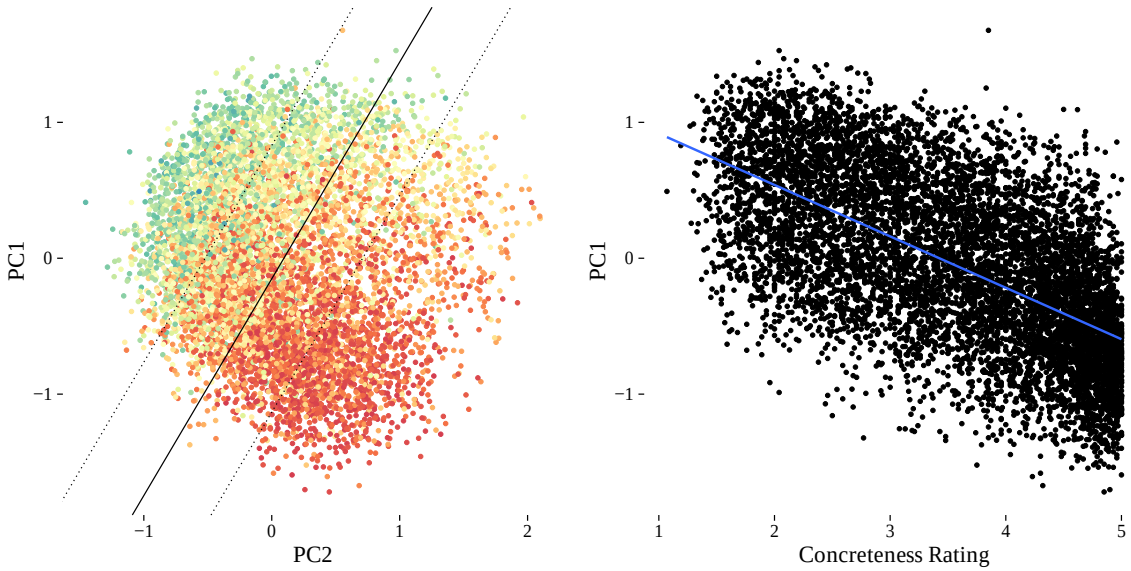


Figure 5.13 (a) Two-dimensional projection of the word embeddings for the nouns used in the Brysbaert et al. (2013) study. The dimensionality reduction was done using Principal Component Analysis retaining only the first two principal components. Red denotes nouns which have been rated as more concrete by human raters and blue as more abstract. Linear separation was done by fitting a Support Vector Regression model using a linear kernel. (b) Correlation between the first principal component and concreteness ratings from the Brysbaert et al. (2013) dataset. There is a clear negatively monotonic relation between the two variables $\rho = -.62$, $p = 0$.

does not need to be considered as measuring abstractness[13] rather the indirectly related fact that abstract words have different relations to other concepts than concrete ones.

Perceptual characteristics

The discussion in §5.6 points to the direction that there are good reasons to believe that *size* is somehow stored in the cognitive system but that it would probably not be reflected in the distributional representations of words. In this section, we provide direct evidence for both these claims by looking at featural representations (i.e., the McRae norms) showing that they provide qualitatively different information to DSMs. To start with, Table 5.4 shows the

[13]The negative correlation between the first principal component and concreteness from the Brysbaert et al. (2013) ratings implies that pc1 is related to abstractness rather than concreteness. However, as noted above, the sign of the principal components identified by PCA is random so an equivalent solution would be one that flipped the sign of pc1 rendering its interpretation as a concreteness estimate.

Table 5.4 The ten most frequent features given as responses in the McRae norms (McRae et al., 2005).

Feature	Frequency
`made of metal`	0.25 (133)
`is small`	0.22 (121)
`is large`	0.20 (106)
`an animal`	0.18 (99)
`is long`	0.15 (81)
`made of wood`	0.15 (79)
`is edible`	0.14 (78)
`is round`	0.14 (76)
`different colours`	0.11 (58)
`is brown`	0.10 (54)

Note: The numbers on the left denote the proportion of nouns containing that feature whereas those in the parentheses indicate the number of occurrences of each feature.

first ten features that appear in the McRae norms (McRae et al., 2005) regarding occurrence. Interestingly, the table mainly contains features one would characterise as 'perceptual' as they refer mostly to physical attributes of the concept rather than semantic relations.

We further establish this dissociation between *experiential* and *distributional* information by looking at featural representations of a set of words which have been marked either as `is_large` or `is_small` and the distributional representations of the same set of words. In the McRae norms, 217 (ca. 40%) words possess size information of which 101 were large and 117 small. For these words, we extract their features ($N = 1225$), and then we construct binary vectors as in §§ 3.3.2 and 5.3.4. The resulting vectors were quite sparse (on average, about 1% of each vector was non-zero) and were subsequently subjected to our dimensionality reduction algorithm to generate the two-dimensional representation shown in Fig. 5.14a. We see that despite some minor errors large concepts cluster differently than small ones. Note that there are clusters in Fig. 5.14a (one larger on the left, one on the right, and one on the bottom) and recall that the t-SNE algorithm preserves the local structure of the high-dimensional space, so we look for the differences in each neighbourhood instead of differences along an axis. Conversely, the distributional representations of the same words do not carry any information regarding their size as shown in Fig. 5.14c. While certain sub-clusters are preserved in both representations (e.g., *vehicles* or *types of ammunition*), the feature norms further demarcate for size whereas the neural embeddings do not.

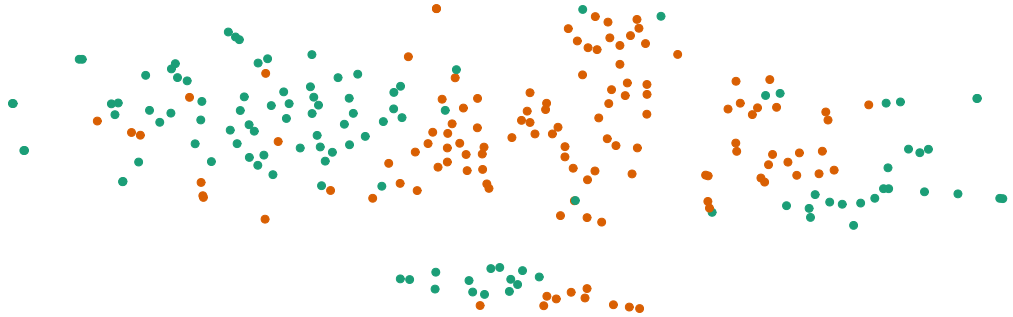
Following Rosch (1978), however, the features spanned by each concept exhibit correlational structure (e.g., [+HAS WINGS] correlates with [+CAN FLY]). It might, therefore, be the case that it is not *size* which pulls the concepts apart in Fig. 5.14a, but an interaction of other features. This finding could be significant in the present case, as it would tell us that we need not encode size information in the semantic representations directly. To this end, we repeat the same procedure as above omitting the relevant feature of *size*. That is, we omit the `is_large` and `is_small` dimensions. Figure 5.14b presents the two-dimensional projection of the vectors without the relevant feature. Admittedly, the results now look a lot more like the distributional vectors than the experiential features suggesting that at least for those words size information is not recoverable by other means.

The above results are congruent with results from semantic priming of perceptually similar items. Table 3.5 shows that in the Semantic Priming Project, perceptual properties do not exert any priming effects, a result which is also predicted by the neural embeddings model. Early work on semantic priming (Flores d'Arcais, Schreuder & Glazenborg, 1985; Schreuder, Flores d'Arcais & Glazenborg, 1984) has shown that semantic priming based on the perceptual characteristics of the stimuli is possible. However, their results have been questioned by Pecher, Zeelenberg & Raaijmakers (1998) who argue that any priming effects were due to a combination of longer SOAs, conscious strategies, and repetition effects (for a summary of their criticism, see Hutchison, 2003a). Pecher et al. (1998) show that when controlling for all these nuisance factors, any priming effects disappear and can only appear when the participants are pre-trained to notice the relationship between the prime and the target (i.e., form a conscious strategy). These results not only support our above argument that DSMs provide the best description of semantic priming effects, but they also highlight the close relationship between semantic priming and tasks of semantic implicit learning.

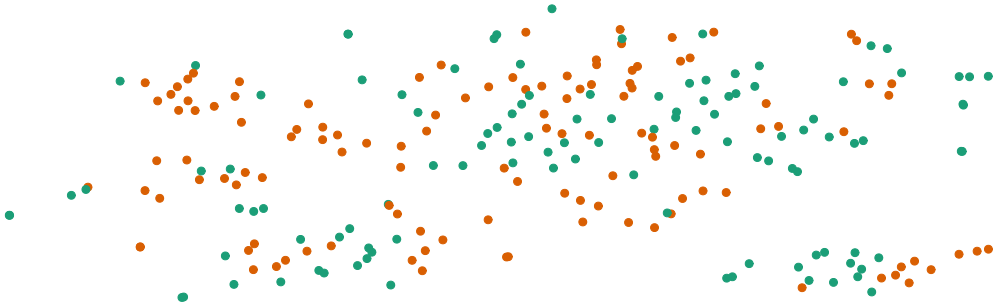
5.8.3 Computational considerations

Architecture of the model

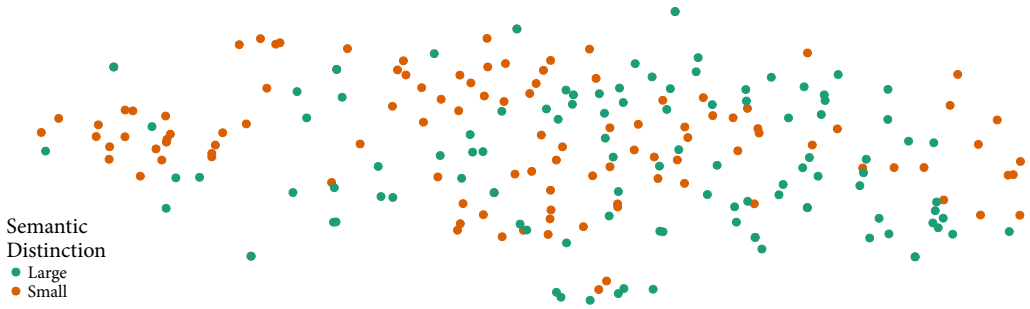
The architecture and learning algorithm of the network raise a potentially important modelling issue regarding *supervised* vs. *unsupervised* approaches. The models used throughout the present study are *supervised* in the sense that there is a known gold output the value of which the model is trying to match given a particular input. Intuitively, however, implicit learning is an unsupervised form of learning, where the learner picks up regularities from her environment *incidentally* (i.e., without any intention to do so). The current architecture is then, perhaps, more suited to problems where the participant knows that there is an output that she needs to emulate and our interest as modellers is to evaluate the generalisation patterns



(a) Representations from the McRae norms



(b) Representations from the McRae norms without size information



(c) Distributional representations

Figure 5.14 Two-dimensional projection of 217 words which include either `is_large` (N = 101) or `is_small` (N = 116) as features in the McRae norms (McRae et al., 2005). (a) Word representations based on the McRae feature norms (for the derivation see text). The t-SNE algorithm shows that *small* concepts cluster differently than large ones (for an explanation on the clusters see text). Classifying by size achieves an error rate of 0.29 with a linear support vector machine classifier. (b) Word representations based on the McRae norms having excluded the two relevant to size features (`is_large` and `is_small`). Classification error was 0.44. (c) `word2vec` distributional representations. Regarding classification error, the values were similar to the McRae feature norms *without* the size information at 0.44.

given the training input. An example of such a problem which can be solved by a supervised learning model is learning the past tense of English verbs as described in §1.3. The learner is provided with input-output pairs corresponding to verbs and their past tenses. For both the learner and the model the task is to associate the two forms and for the modeller to evaluate whether the performance on unseen instances agrees with the behavioural data.

The above discussion prompts us to consider how would we construct an unsupervised model of SIL. We have already presented a simple solution at the start of each simulation in the form of a two-dimensional projection of the stimuli used in each experiment. Most dimensionality reduction algorithms, apart from being helpful visualisation tools, are *unsupervised* learning models that seek to compress high-dimensional data highlighting important aspects of it and reducing random noise. This idea is closely related to *manifold learning* in machine learning which asserts that in high-dimensional datasets most dimensions are redundant containing mainly noise. According to this, we need to look at only very few dimensions to achieve considerable levels of generalisation as only these provide relevant information (see also, Shepard, 1980, and Shepard, 1987, for applications in psychology). We have already introduced PCA, MDS, and *t*-SNE as methods for dimensionality reduction; PCA works by selecting the dimensions that capture the most variance, MDS by preserving the global structure of the high-dimensional dataset, whereas *t*-SNE focuses on the local structure. Another method closely related to the neural network described above would be a self-organising map (Kohonen, 1982), which also projects the input on a low-dimensional manifold (commonly a two-dimensional grid). This method is particularly attractive when considering classification problems, as projecting the high-dimensional input on a two-dimensional lattice gives us the area spanned by each representation making classification easier. All these approaches have in common is that they project their input to some lower dimensional space and generalise by measuring the distance between each novel item and the aggregate of each group.

While attractive this approach is not without problems; Take Fig. 5.2a as an example. We see that the two semantic categories are neatly divided between animates and inanimates yielding much higher within- than between- cluster similarity. In other words, each word used during the testing phase is closer to its grammatical alternative than its ungrammatical. For the sake of argument consider a setting similar to the behavioural study in Williams (2005), where the participants see the nouns paired with novel determiners. Consider now that in the 2AFC task, the probability of choosing the correct determiner is equal to the distance between the target and the centroid of each determiner cluster. In other words, in the case of $\left[\frac{\text{gi}}{\text{ro}}\right]$ 'dog', where 'dog' was previously seen with the determiner *ul*, the probability of choosing *gi* instead of *ro* is equal to the distance between 'dog' and all the words that went with *gi* during training vs the things that went with *ro*. Undoubtedly, because of the high within cluster

similarity, the probability of choosing the incorrect determiner is smaller than that of selecting the grammatical alternative.

These problems extend to the related approach of modelling SIL tasks using models of associative memory. Under this account, SIL occurs as a consequence of the processes of long-term memory. Since we have already argued that we can think of the distributional representations as long-term memory traces of the words (Kintsch & Mangalath, 2011), SIL would arise in this paradigm from the way words are encoded, recalled, and integrated with new experiences (see also, Thiessen & Pavlik, 2012). Chubala et al. (2016) model a semantic implicit learning task reported by Neil & Higham (2012) using the MINERVA (Hintzman, 1986) model of associative memory. The main idea is that during the training phase, the long-term memory traces of the words are activated and brought forward forming a composite representation for each group. Concretely, encountering words such as *dog*, *cat*, and *monkey* with the determiner *gi*, results in the *gi* representation being the aggregate of its components. During testing, the participants are assumed to select the representation that minimises the distance with the probed noun. However, this aggregation would in the limit face the same problems as the dimensionality reduction techniques above, as the determiner group centroid will always be much closer to the grammatical alternative than the ungrammatical one.

We solve the above problems by further asserting that the fact that learning is unconscious and incidental does not necessarily mean that there are no teaching patterns (see also O'Reilly & Munakata, 2000). Consider the model introduced in Fig. 2.2 for example; in some sense, this should be an unsupervised model as we do not include any other information apart from the words in the corpus. However, the algorithm outlined in §2.3.1 casts the problem as a supervised learning one by treating the context of each word as the teaching pattern used to induce the semantic representation. In the case of SIL tasks, while we do not provide participants with input → output patterns overtly, we assume that they can internally transform this input by selecting the relevant words (the novel determiner and the noun) and cast the problem as a supervised learning one (associate the noun with the corresponding determiner). Defining the problem this way enables the participants to learn in an error-driven manner as the backpropagating neural network (see, e.g., Fine & Jaeger, 2013).

Point Estimates

We identify a potential problem in the method outlined in §5.3.2 to extract point estimates from the generalisation gradients resulting in different numbers of epochs in different studies. More specifically, our objective in §5.3.2 was to maximise the fit between the network estimates and the reported performance. However, our results might be biased in the sense that we cannot necessarily equate the network's knowledge at the epoch which maximises the fit

with the knowledge of the learners at the start of the test phase. We recognise the validity of this counterargument, noting, however, that we mainly draw our conclusions from the overall generalisation patterns of the network instead of the point estimates. After all, the model provides a computational abstraction which we evaluate empirically, instead of an oracle. In any case, we briefly outline a method that could be explored in future research and tackles this issue. The problem identified above was that instead of selecting the point estimate based on the performance on the training set we select it by maximising the fit to the test set. Selecting the point estimate based on the training set *should* be a better alternative as it provides an objective baseline figure. A solution, therefore, we propose would be to incorporate this knowledge (the performance on the baseline) as a prior term in (5.6). This Bayesian solution would maximise the fit to the data, on the one hand, modulating this fit by the probability of the epoch, where the prior probability for each epoch could be sampled from a distribution centred around the reported value. This way we would discard any epochs at which the network is either over- or under- trained increasing our confidence in these estimates.

Chapter 6

Checking the assumptions

6.1 Introduction

This chapter contains two studies that seek to verify two claims we have taken for granted so far. Firstly, that surface-level phonological information cannot explain the results of the experiments (see §§ 3.1 and 5.1). If that were the case, then phonological information would provide a simpler explanatory model for the data obtained in the behavioural experiments presented in §1.5. The reason for assuming that this model is simpler is because in this case, participants would not need to activate *any* semantic representations, rather simply associate statistical patterns present in their current input. Secondly, we argue in §5.2 that participants rely on the *backwards* probabilities in the sentence to learn the association between the novel determiner and the noun. This behaviour might seem counter-intuitive considering the order of the elements in the sentence. The study illustrated in §6.3 show that across languages *backwards* probabilities can be just as informative for the learner, but also qualitatively different.

6.2 Study 9: Is Semantic Implicit Learning really semantic?

6.2.1 Introduction

The generalisation patterns illustrated in the above experiments depend on whether unconscious extraction of regularities based on semantic category membership is possible. Experimental data using the *contextual cueing* paradigm (Goujon, 2011; Goujon, Didierjean & Marmèche, 2009; Goujon, Didierjean & Thorpe, 2015) as well as studies on *semantic priming* (see Chapter 3) suggest that there is an availability of semantic (or, at least, semantic-like) information even when it is not necessary in order to carry out the task. Goujon et al. (2009), for instance, presented participants with a search task displaying words in random positions

on a screen where participants had to indicate whether a target word appeared on the left or right side of the screen. Crucially, the semantic class of the contextual words depended on the position in which the target words appeared on the screen. The semantic classes used in these experiments were more specific than the semantic distinctions in the experiments outlined in §1.5. The categories used were mammals, birds, trees or flowers, and fruits or vegetables.[1] The authors found that there was a significant speedup for the trials where semantics was predictive of the position compared to a random baseline suggesting that the participants used the semantic context provided by the other words. Further experiments in which a natural scene instead of an array of words provided the context corroborated these results (Goujon, 2011).

These results support the premise of the studies presented in §1.5. However, it remains an open question as to whether we can equate semantic implicit learning in tasks using natural language stimuli with those involving visual search. In what follows we argue that the linguistic component of the above tasks renders them markedly different from the tasks of contextual cueing, introducing a variety of potential confounds.

In the experiments presented above, participants learn a grammatical system akin to the gendered systems found in many of world's languages where a small set of determiners is used to classify large noun categories. In such languages, classification is thought to be arbitrary (Corbett, 1991) (also Ex. 6.1–6.4) However, large-scale studies in German have uncovered that morpho-phonological factors seem to be consistent predictors of a word's gender (Petig, Hammer & Durrell, 1993; Zubin & Köpcke, 1984). For example, words ending in *-er* are masculine, whereas words ending in *-ung* are feminine. The same studies have shown that apart from a handful of exceptions (such as names of seasons, months and days of the week which are canonically masculine), semantic distinctions are not predictive of gender class. Could it be the case then that the phonological features of the experimental stimuli drive the generalisation patterns and not the semantics of the words? After all, in the post-experimental interviews conducted to assess awareness of the relevant variables (e.g., animacy) participants mention the phonology of the noun as their basis for generalisation (Williams, 2005, p. 284). Further to this, knowledge of a gendered language also correlated, albeit mildly, with generalisation performance (Williams, 2005, p. 288). While the experimenters in such tasks have taken care to eliminate obvious confound factors such as salient phonological cues (e.g., all words in one category end with the same phoneme), in what remains we will examine further whether more fine-grained phonological features can explain the behavioural performance in the experiments.

[1] In subsequent experiments the authors increased the number of categories (e.g., including fish) but keeping constant the level of specificity.

(6.1) *Die* *Frau*
ART.1SG.**FEM** woman

(6.2) *Das* *Weib*
ART.1SG.**NEU** woman

(6.3) *Das* *Auto*
ART.1SG.**NEU** car

(6.4) *Der* *Wagen*
ART.1SG.**MASC** car

Why would participants be more inclined to rely on phonological cues than semantics in these tasks? Could it be because humans process phonological information before semantics? An array of recent neurolinguistic studies (Hauk, Davis, Ford, Pulvermüller & Marslen-Wilson, 2006; Miozzo, Pulvermüller & Hauk, 2015) using a combination of neuroimaging methods, regression analyses and a variety of lexical decision tasks, have shown that during the time-course of word reading phonological information is activated simultaneously with semantic variables as early as 150ms after the presentation. Also, influential computational models of word reading (Plaut, 1997; Plaut et al., 1996) consider that semantic information activates with phonological before a word is fully recognised. The results of these studies capitalise on the fact that it is not the case that phonological information takes precedence in lexical processing pushing semantic information to the periphery.

The answer to the above question might be given in experiments such as those reported by Frigo & McDonald (1998) and Brooks, Braine, Catalano & Brody (1993) which explore the learnability of gender-like subclasses. In these experiments, Frigo & McDonald (1998), for example, constructed artificial languages overtly marking gender using phonological cues. They then manipulated several factors attempting to find what conditions would enable participants to achieve higher generalisation rates. They found that *perceptual salience*, *position* and *frequency* of the phonological markers facilitated the generalisation of the respective gender subclasses. However, contrary to the tasks at hand, one major component of these experiments was to draw attention to the link between the indicator (here the determiner) and the phonological marker.

Modern linguistics has vehemently dismissed the idea of *sound symbolism* (de Saussure, 1916) from its early days as it assumes that each word gains its meaning by relation to other words and not its phonological representation which is thought to be arbitrary. However, large-scale corpus studies have shown a mild systematicity between form and meaning; regarding grammatical category learning, Monaghan, Chater & Christiansen (2005) have shown that a small set of phonological features (Table 6.2) can be used to distinguish between open- and

closed- class words and between nouns and verbs. Additionally, more recently, Monaghan, Lupyan & Christiansen (2014a) have shown that semantic attributes such as *sweet* or *liquid* have phonological correlates and allow for semantic categorisation of non-words. Results from research in phonology (Hinton, Nichols & Ohala, 1994) claiming that, albeit to a limited extent, certain semantic attributes are manifested similarly in languages around the world, citing evolutionary advantages of this further corroborate the above. Furthermore, §A.7 outlines a study that explores whether it is possible to classify animacy and concreteness only from phonological features. Using WordNet we gather nouns falling in either category which we then use to extract the phonological features described below. Although the effects are quite small, we can achieve better-than-chance semantic classification based solely on phonological features. While for the purposes of the present study we do not need to subscribe to either position, our simulations implcitly suggest that it is possible for semantic distinctions to find correlates in phonological representations.

In light of the above results, we explore two possible ways in which phonology might explain the results of the above studies. Firstly, the stimuli used in the experiments outlined in §1.5 might differ in their phonological representations biasing the participants to classify the nouns in a certain way. For example, while the researchers have taken care to eliminate obvious cues (e.g., 'gi' words start with /b/), participants could still form hypotheses resulting from rehearsal in their phonological memory (e.g., Baddeley, 2003) such as '*"gi" words tend to be longer*'. Secondly, even if the stimuli used do not differ in their characteristics it might be the case that some combination of these features makes them learnable. We explore the former hypothesis by comparing the phonological features on a by category basis, while for the latter we use a learning model similar to §5.3.1 to simulate the tasks using phonological, instead of semantic, representations.

The limited amount of experimental stimuli together with the lack of overt phonological markers renders the first hypothesis less probable unless the cues are particularly salient. The second hypothesis, however, can provide a simple model of the results at hand. While it remains an open question as to whether semantic distinctions are manifested phonologically in general, here we test whether a simple set of phonological features, similar to the one used in Monaghan et al. (2005), can be used to classify the nouns used in the above experiments. The implication of this small study is that if a simpler, phonologically-motived simulation, could explain the patterns observed, Occam's razor would have it that this should have been selected.

6.2.2 Materials and methods

We derive phonological feature vectors for the experimental stimuli used in Williams (2005) and Paciorek & Williams (2015) using the Carnegie Mellon Pronunciation dictionary[2] to extract the phonological information for each word and the TEX hyphenation algorithm described in Liang (1983) for syllabification. The `cmudict` contains the pronunciation of 133287 English words in ARPABET format. For example, it represents the word *Cambridge* as `K EY1 M B R IH0 JH` which is equivalent to the International Phonetic Alphabet (IPA) /ˈk eɪ m b ɹ ɪ ʤ/. A number after a vowel denotes either primary or secondary stress. The 1 after the diphthong `EY` denotes primary stress and the 0 after `IH` that there is no stress (Cambridge does not have secondary stress which would have been marked by a 2 after the corresponding vowel). Using the above scheme, we derive phonological feature vectors based on Monaghan et al. (2005). The values used for *vowel position* and *vowel height* were derived by evenly splitting the vowel space as shown in Table B.2. In Table 6.1 we show some examples of the corresponding phonological representations.

6.2.3 Results

We first ask whether the stimuli groups used in the behavioural experiments differed significantly regarding their phonological features. Since in all the experiments the participants would only see the two levels of *one* semantic distinction we compare only those two levels (i.e., abstract vs. concrete and animate vs. inanimate not, e.g., abstract vs. animate). Three features were highly correlated ($|\rho| > 0.5$) in our dataset, *phoneme* and *syllable length* ($\rho = 0.8, p < 0.001$), *vowel* and *stressed vowel position* ($\rho = 0.68, p < 0.001$) and *onset* and *word complexity* ($\rho = 0.53, p < 0.001$). We firstly look at the dataset of Williams (2005); the inanimate nouns had significant greater *phoneme length* ($t(25.59) = -2.47, p = .021, d = -0.87$), while the animate ones were more likely to have a higher proportion of *nasals* ($t(18.45) = 2.29, p = .034, d = 0.81$). Turning to the high similarity dataset in Paciorek & Williams (2015), there were no significant differences between abstract and concrete words regarding their phonological features. Interestingly, in the low similarity experiment there were more significant differences between the features; the abstract words had greater *phoneme length* ($t(38.65) = 3.06, p = .004, d = 0.88$), *syllable length* ($t(37.39) = 3.16, p = .003, d = 0.91$), and *position of stress* ($t(23.00) = 4.03, p < .001, d = 1.16$) (all of which are related), while the concrete words were more likely to have a *reduced first syllable* ($t(45.82) = -2.41, p = .02, d = -0.69$). The significant differences in the low similarity experiment are somewhat counter-intuitive as there were no significant differences between the two groups in the behavioural

[2] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

Table 6.1 Example representations using the phonological features used in Monaghan et al. (2005) to distinguish between different grammatical categories. For a description of each feature and how it was computed see §B.4.

	TABLE	OXYGEN	TURTLE
Word level			
Length in phonemes	5.00	7.00	5.00
Length in syllables	1.00	2.00	1.00
Presence of stress	1.00	1.00	1.00
Position of stress	1.00	1.00	1.00
Syllable level			
Onset complexity	1.00	0.00	1.00
Word complexity	0.60	0.57	0.60
Proportion of reduced vowels	0.00	0.00	0.00
Reduced first syllable	1.00	0.00	1.00
Phoneme level			
Coronals	0.67	0.50	1.00
Initial /ð/	0.00	0.00	0.00
Final voicing	3.00	3.00	3.00
Nasals	0.00	0.14	0.00
Stressed vowel position	0.25	2.00	1.00
Vowel position	1.12	2.00	1.50
Vowel height	1.38	2.33	2.00

Note: The original featureset used by Monaghan et al. (2005) also includes a feature marking *-ed* inflection; however, since this feature would only be meaningful in distinguishing between adjectives and verbs (Kelly, 1992) we did not include it in any analyses.

experiment. Fig. 6.1 shows a two-dimensional projection of the phonological space using *Principal Components Analysis* revealing no apparent clusterings between semantic categories.

The significant differences between the features of the stimuli used in the experiments, prompt us to explore further whether the grammatical systems introduced in §1.5 are recoverable solely on phonological grounds. We similarly test this hypothesis as we have done before by treating this as a multiclass classification problem. Since the number of predictor variables (i.e., the phonological features) is much smaller than in the case of semantics, where we technically had 300 predictor variables, we do not need the computational capacity of a neural network in the present situation, so we opted for using a simple logistic regression model. As in Chapter 5, we feed the training matrix $\mathbf{X}$ along with a response vector for each of the determiners $\mathbf{y}$ to the classifier and then record the returned probabilities for each of

Table 6.2 By-feature means for the stimuli used in the semantic implicit learning experiments presented in §1.5 using the phonological cues derived in Monaghan et al. (2005).

	W05		PW2015–H		PW2015–L	
	Animate	Inanimate	Abstract	Concrete	Abstract	Concrete
Word level						
Length in phonemes	3.25	4.31	7.25	6.46	4.62	6.21
Length in syllables	1.12	1.25	2.33	2.04	1.38	2.00
Position of stress	1.00	1.00	1.25	1.50	1.00	1.54
Syllable level						
Onset complexity	1.12	1.31	0.92	1.08	1.21	0.92
Word complexity	0.61	0.66	0.59	0.63	0.66	0.63
Prop. of reduced vowels	0.44	0.12	0.03	0.01	0.00	0.00
Reduced first syllable	0.88	0.81	0.17	0.42	0.62	0.29
Phoneme level						
Coronals	0.39	0.50	0.69	0.64	0.58	0.56
Final voicing	0.50	0.62	0.50	0.96	0.71	0.58
Nasals	0.12	0.02	0.17	0.13	0.10	0.11
Stressed vowel position	0.81	1.16	0.77	0.68	0.84	0.48
Vowel position	1.05	1.38	1.11	1.21	1.22	0.93
Vowel height	1.39	1.30	1.42	1.47	1.44	1.46

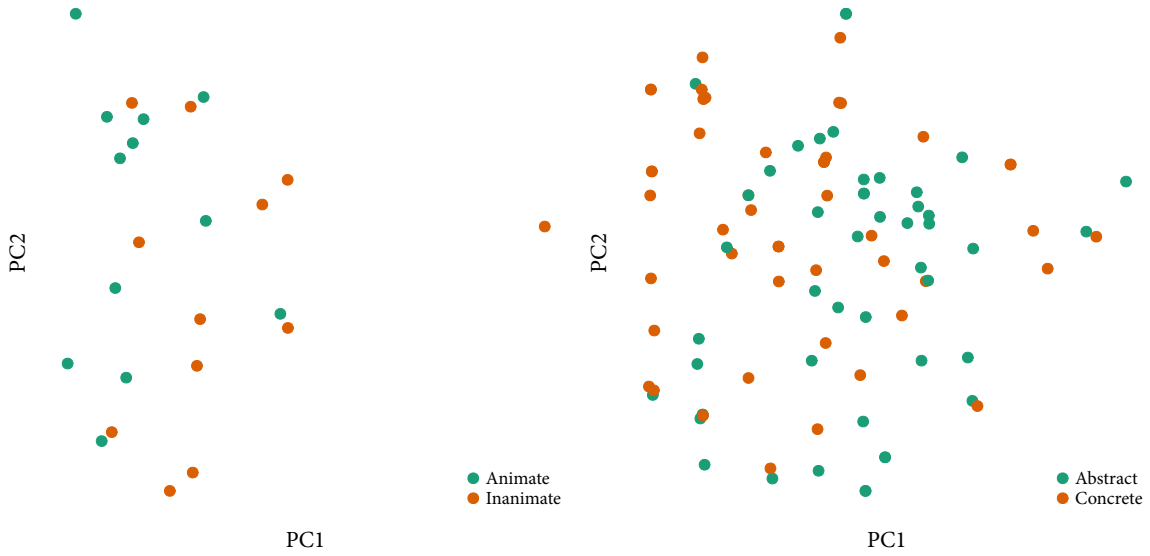
Note: W05 = Williams (2005); PW2015–H = Paciorek & Williams (2015, High similarity); PW2015–L = Paciorek & Williams (2015, Low similarity). We drop the features 'Presence of stress' and 'Initial /ð/' from any comparison or simulation that was done on the Williams (2005) and Paciorek & Williams (2015) datasets as their values remained constant.

the stimuli in the test set. Since the logistic classifier is deterministic (i.e., returning the same results every time), we simulate different 'learners' by adding random noise to the training data. This noise represents potential issues in encoding, processing or other possible factors that might affect performance. Note that we add the noise on a per-learner basis in that no two learners 'experience' their input in the same way.

$$\mathbf{X} + \mathbf{R} \sim \mathcal{N}(0, 1) \in \mathbb{R}^{N \times D} \tag{6.5}$$

where $\mathbf{X}$ is the design matrix, $\mathbf{R}$ the random noise matrix, N the number of samples, and D the dimensionality of the feature matrix (here, 15).

Having set up the classifier and the input data, we simulate the behaviour of 30 learners averaging by participant. For the animate vs. inanimate distinction (Williams, 2005), the



(a) Animate – Inanimate **(b)** Abstract – Concrete

Figure 6.1 Two-dimensional projection of the phonological space of the nouns used in the experiments done by Williams (2005) and Paciorek & Williams (2015).

performance of the classifier was not significantly better than chance contrary to the human experiments ($t(29) = -1.27, p = .214, d = -0.23$). Looking at the endorsement rates in the high similarity condition of the concreteness dataset, the classifier predicts a null effect with a slight numerical advantage for the ungrammaticals ($t(29) = -0.43, p = 0.67, d = -0.08$). Similarly, in the low similarity condition, no effect is observed ($t(29) = 0.45, p = 0.66, d = 0.08$). Analysing the task as a 2AFC task, that is, instead of comparing the two conditions, see if the classifier is selecting the grammatical alternative more often than chance, the results become even more mixed; for the high similarity dataset the classifier predicted a significantly lower than chance performance ($M_{\text{High Similarity}} = 0.44, t(29) = -3.38, p = .002, d = -0.62$). Conversely, the simulated 'learners' performed better than chance in the low similarity condition ($M_{\text{Low Similarity}} = 0.57, t(29) = 4.31, p < .001, d = 0.79$).

6.2.4 Discussion

In the above simulations, we test the hypothesis that surface phonological features can also explain the generalisation gradients observed in the semantic implicit learning experiments outlined in §1.5. We tested this hypothesis by deriving 15 phonological features using a scheme used elsewhere in the literature (Monaghan et al., 2005) for all the words in the experiments.

We then ask whether some features were more predictive for some groups, and using a simpler version of the model in §5.3.1 whether a simulated learner would be able to generalise as a real learner. The results did not support this hypothesis; in none of the experiments the simulated results agreed with the behavioural ones. Concretely, for one of the experiments the model did not predict the observed higher than chance rates, whereas, for the rest of the experiments, the model either predicts effects in the opposite direction or patterns not found in the behavioural results. Taken together these results suggest that, at least for the present datasets, phonological cues did not inform the participants during the task.

The fact that the underlying grammatical systems of the studies presented in §1.5 are not recoverable by phonological cues does not imply that *any* such system cannot be learnt from phonological information alone. When experimenters control for such cues, phonological patterns (Frigo & McDonald, 1998; Lany & Saffran, 2010; Monaghan et al., 2005) aid in word segmentation, acquisition of word categories, and subsequent semantic integration (but also, see a notable exception in Ouyang et al., 2016 and the discussion therein). In fact, the simulations show that relying solely on phonological information, the most discriminative system (the low semantic similarity one) *should* be learnable, whereas the opposite should hold for the less discriminative one. This pattern is not, however, one we observe in the behavioural data, where learners rely on the distributional patterns of words instead of their phonological characteristics. It follows that either phonological information is not crucial at that stage or the way the task is set it biases learners away from these cues. In any case, the learners do not *use*[3] phonological information as a predictor. On the other hand, such cues might play an adversary role in these experiments as they could add more noise to the participants' emerging knowledge. Future work could, therefore, explore this interaction between phonological and semantic cues and the participants' reliance on them during the tasks.

A potential drawback of the above simulations would be that the phonological features used here were insufficient to carry out the classification. In other words, it might be the case that participants *do* rely on phonological cues, but not on these, or even a reduced version of those as implied by the PCA figures (Fig. 6.1). Indeed, for the semantic representations we performed a rigorous examination (Chapters 3 and 4) of which description would fit the results better. While this is a fair counterargument, we note that the number of possible phonological feature vectors that we could construct is much smaller than that of semantics. As an alternative, we could have followed Monaghan, Shillcock, Christiansen & Kirby (2014b), Monaghan, Christiansen, Farmer & Fitneva (2012), and Harm & Seidenberg (1999) who use

[3]We note an important distinction at this stage between whether participants consider some cue to be important in the experiment and whether they *use* it predictively. Post-hoc experimental questionnaires show that participants sometimes consider phonological cues as relevant. However, their performance cannot be explained solely on this basis.

only phoneme-related features (e.g., /k/ is velar, plosive, etc.). However, while this would be a valid alternative, it readily works only for similarly sized words, which is not the case here. The problem is that any classifier[4] would require fixed-size input representations which we would not be able to obtain by simply concatenating the features of the experimental stimuli.

To our knowledge apart from some accounts in animistic societies (Nuckolls, 2010), there is no single featureset that predicts animacy or concreteness in English (apart from some subcategorisations of them, e.g., Monaghan et al., 2014a). It might be the case that such features exist but cannot be studied by segmental phonology. Indeed, suprasegmental cues are attended to by language learners (Mehler, Bertoncini, Dupoux & Pallier, 1996) and could potentially aid participants in such tasks. Alternatively, derivation of meaningful features directly from the speech signal (Kiela & Clark, 2015) might be more informative for the classifier. While further computational results might shed more light on whether 'core' semantic distinctions such as those used to construct the grammars in the implicit learning experiments above need to be done, the above results suggset that in the given datasets category prediction solely by phonological features is not possible.

6.3 Study 10: Comparing forward and backward probabilities

6.3.1 Introduction

The computational model we built in §5.2 associates the distributional representations of nouns with particular determiners. We assumed that this is a reasonable abstraction of the tasks, so long as participants can (a) attend to the relevant forms, and (b) use a retrodictive (that is, classification) mechanism instead of a predictive one. Through a brief linguistic analysis offered in §5.2, as well as because of the structure of the experiments, we can assume that participants attend to the relevant forms (the noun and the determiner). On the other hand, assuming that participants use retrodiction instead of prediction is a more contentious issue. Recall that retrodiction is assessing the probability of a word *preceding* another, whereas in prediction we evaluate the probability of a word following another. For example, consider the German phrase 'der Mann'; prediction here would not be as informative as the set of words that can *follow* 'der' is much larger than the set of words that can *precede* 'Mann'. Retrodiction, as a statistical cue, has been found to influence participants' experience of experimental stimuli in a variety of occasions (Jones & Pashler, 2007; Onnis & Thiessen, 2013; Pelucchi et al., 2009; Perruchet & Desaulty, 2008).

[4]This problem could be mitigated by building a Recurrent Neural Network §§ 1.3 and 2.3.2 that takes into account the temporal dependencies between the phonemes. This alternative might be better suited for a more rigorous analysis in future work.

Our discussion has so far has assumed that this probabilistic structure of languages around the world is particularly important for the learners both in the case of L1 (Thiessen & Saffran, 2007) and L2 (Onnis & Thiessen, 2013). We have already sketched a toy model of word segmentation (§1.2) that takes into account the probability of a syllable given the ones that preceded it. In fact, corpus analyses performed by Swingley (1999) showed that word segmentation could be achieved not only by looking at the forward transitional probabilities of syllables but also by looking at the backwards. Furthermore, current state-of-the-art language models used in speech recognition (Graves et al., 2013) regularly use both forward and backwards passes over the same sequence to achieve better performance.

On the other hand, while statistical cues are ubiquitous in the languages around the world, it is not the case that all the languages share the same statistical structure. Languages differ in the way they convey information, the number of tokens they need to express similar ideas or the general predictability of any word in the text (Bentz, Alikaniotis, Cysouw & Ferrer-i-Cancho, 2017a). Based on our observation above about the asymmetry in forward and backwards probabilities in ARTICLE → NOUN bigrams, we conjecture that the average information content conveyed by each word in the text conditioned on the preceding words would differ from when conditioned on the succeeding words. We test this hypothesis by looking at the forward and backwards *entropy rates* for ten languages with different characteristics (see Table 6.3).

Intuitively, the entropy of a random system is a measure of its predictability. For example, a fair coin toss has an equal chance of landing heads and tails. Our uncertainty, therefore, is maximised at every flip since each toss event is independent of the ones preceding it. Similarly, we can measure the uncertainty associated with each language by counting the number of times each word appears in a text. The implication of this is that if the words in a language are more equiprobable our uncertainty would be maximised, but moderated in a language where a few words appear more times. If all things being equal, one language that has more word types that appear fewer times is going to be less predictable than a language where fewer word types appear more. This notion, however, assumes that words are drawn from a multinomial distribution according to their frequencies irrespectively of the contexts in which they appear.

While entropic measures are interesting and pervasive in quantitative linguistics (Baayen, 2001), it is easy to see that each word in the language is not independent of the ones preceding it. Recall the example at the beginning of Chapter 2; 'Our pockets were full of …'. Semantic restrictions aside, if we take into account the preceding words, our uncertainty should be less in what would follow. The entropy rate of a random system measures exactly that; our uncertainty of an upcoming symbol given the ones we have seen so far in the sequence. The

entropy rate h of a text T is, therefore, defined as,

$$h(T) = \lim_{N \to \infty} H(t_N | t_1, t_2, \ldots, t_{N-1}) \tag{6.6}$$

where N is the size of the sequence, H the entropy, and t_i the tokens in the sequence.

Furthermore, if languages differ in the ways they encode information, it could be the case that some relations are more predictable in some languages but less so in others. Returning to the ARTICLE → NOUN relations, speakers of languages that do not possess articles might find it harder to learn these constructions as they have different expectancies for the dependencies within the linguistic sequence. From this observation, we conjecture that in languages with articles the backwards probabilities would be more informative, hence more useful to the learners. Based on these arguments we might be able to explain the individual differences observed in SIL tasks (Williams, 2005), where participants show language-specific biases.

We argue here that the participants' use of retrodiction instead of prediction stems from an $l1$ statistical bias in which for some grammatical constructions, retrodiction is more informative than prediction. Onnis & Thiessen (2013), for example, presented adult speakers of Korean and English with auditory linguistic sequences where forward and backwards cues were equiprobable. They found that participants exhibited a strong L1 bias as speakers of English tend to prefer backwards probabilities as their cue, whereas Korean forward. The authors argue that this bias stems from the difference in the head directionality of the two languages; English, being a head-first language, has lower backwards probabilities in the case of preposition → noun bigrams, whereas Korean, being a head-final language exhibits the opposite pattern. In the present case, observing $l1$ influence during implicit learning would not only be informative for subsequent research assessing the learning gradients of speakers of different languages but would also explain the diverging patterns observed in Williams (2005, p. 295) where participants who spoke languages with grammatical gender performed better in the SIL task.

If experience with language alters the way learners utilise the statistical patterns present in the input, then speakers of different languages might be more or less sensitive to the patterns in the data. In this study, we ask whether such statistical cues as forward and backwards transitional probabilities are found cross-linguistically and whether speakers of different languages are predicted to have similar problems in learning those patterns. We do so by conducting corpus analyses of ten European languages from different language families. Crucially, the languages included in this study vary on whether or not they possess grammatical gender, their use of articles and on various lexical diversity measures.

Table 6.3 Information on the languages used from the EUROPARL corpus. Language family refers to the specific group each language belongs to from either Indo-European (Germanic, Romance, and Slavic) or Finno-Ugric (Finnic). Gender and Article mark whether the language possesses grammatical gender and whether it uses articles. We also record the unigram (see text) entropy H for each language. The number of nouns, tokens, and types (number of unique tokens) are also presented.

Language	Language Family	Gender	Article	# NOUNS	# TOKENS	# TYPES	H
Dutch	Germanic	✓	✓	246.722	997.852	29748	9.57
English	Germanic	–	✓	267.181	996.270	18357	9.37
Estonian	Finnic	–	–	355.365	987.372	79381	12.08
Finnish	Finnic	–	–	369.190	994.348	92659	12.34
French	Romance	✓	✓	251.194	1.040.085	25402	9.6
German	Germanic	✓	✓	177.919	990.880	43245	10.36
Italian	Romance	✓	✓	277.143	1.006.786	30555	10.26
Polish	Slavic	✓	–	345.161	987.009	55822	11.6
Slovak	Slavic	✓	–	310.743	986.957	57464	11.58
Spanish	Romance	✓	✓	242.068	993.402	30581	9.58

Note: H = Shannon's entropy measured in bits.

6.3.2 Materials and methods

Corpora

We study the presence of statistical cues in the texts of *ten* European languages as summarized in Table 6.3. These texts are extracted from the EUROPARL corpus (Koehn, 2005), which contains aligned official translations of the proceedings of the European Parliament in 21 languages. From each text we extract the first one million tokens and perform part-of-speech tagging using TreeTagger (Schmid, 1994).[5] Unfortunately, at the time of writing only ten of the 21 languages in the EUROPARL corpus were supported by the Part of speech (POS)-tagger so we focus solely on them. Subsequently, we preprocess the corpora in a manner similar to the one described in §2.4.2. We transform all the capital letter characters to lower case, we tokenise and remove any punctuation, and, finally, we transform the words in bag-of-words representations (replacing each word by an index). The latter step is done to avoid problems of entropy rate estimator (see below) in texts with non-ascii characters.

[5] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Entropy Rates

We compute entropy rates for the languages presented in Table 6.3 using the algorithm outlined in Bentz et al. (2017a) and Gao, Kontoyiannis & Bienenstock (2008). This estimator applies findings from optimal compression algorithms (Ziv & Lempel, 1978) by taking increasingly larger chunks from the input sequence and counting how many times the shorter sequences re-appear. The average number of matches for all word tokens reflects the redundancy in the input string, or, in other words, the predictability of that sequence. Bentz et al. (2017a) applied this to more than 1000 translations of the Bible and found that there is a strong relation between unigram entropies and entropy rates in that knowing the preceding context reduces uncertainty by roughly the same amount across languages.

Computing the entropy rate when the current word is conditioned on its succeeding tokens is, unfortunately, non-trivial. From a theoretical point of view (6.6) assumes that the entropy rate will approximate the true entropy in the limit (i.e., as the length of the sequence reaches infinity) as the conditions of stationarity and ergodicity are met. In short, stationarity means that the statistical properties of the sequence do not change over time (e.g., as we look at more word tokens), while ergodicity means that the statistical properties of a sufficiently long sequence will match those of the ensemble of all possible sub-sequences. However, while the problem is trivialised when we condition each word on its preceding context, as we can take increasingly large sequences until the two conditions are met, this is non-trivial in the opposite case. We counter this problem by reversing the words in the corpus. For example, a sequence 'Our pockets were full of …' would become '…of full were pockets Our'. While this might seem counter-intuitive note that when we compute the entropy rate at "N = Our" based on its preceding tokens in the reversed sequence, we also compute the entropy rate on the *succeeding* tokens in the original sequence. Considering that the corpus size is large enough for our measures to be stable, reversing it means we can readily plug in (6.6) and measure the average information content in languages that is carried by *any* token given its preceding *and* succeeding tokens. We call the former case the *forward* entropy rate, while the latter the *backwards*.

Transitional probabilities

Together with entropy rates we also measure the transitional probabilities of specific word bigrams in the corpora. Bigram probabilities are intuitively related to the entropy rates as they are proportional to them when $N = 2$. We note, however, that bigram language models, that is, models that predict the upcoming word in a text considering only the previous word exhibit

higher entropy estimates[6] and are quite poor estimators of the overall uncertainty in a language. We extract all bigrams where the second element, or the first in the reversed sequence, is a NOUN. Furthermore, for a random baseline, we also extract an equal amount of random bigrams from the corpus sampling repeatedly without replacement. We compute Forward Transitional Probabilitiess (FTPs) and Backward transitional probabilitiess (BTPs) using (5.2) and (5.3). Since the probabilities can be quite small, we avoid numerical underflow issues by transforming them to an information theoretic measure to examine how much information is conveyed from each relation. We use the *Shannon information content* (Shannon, 1948) or *surprisal* (Hale, 2001) which can be defined as the log of the inverse of the probability of an event.

$$\log \frac{1}{TP} \tag{6.7}$$

where TP in this case can either be FTP or BTP.

6.3.3 Results

We firstly look at the entropy rates of the languages outlined in Table 6.3. Figure 6.2 shows the forward and backwards entropy rates per thousand words in the ten languages from the EUROPARL corpus. We see that the entropies need about 50000 tokens to converge to a stable value. Firstly, the entropy rates seem to follow the unigram entropies in Table 6.3. Bentz et al. (2017a) found that there is a strong underlying linear connection between unigram entropies and entropy rates with an observed difference of about 3.17 bits/word. Secondly, backwards entropies are higher in general than forward entropies, suggesting that it is harder to be certain about the elements that precede a word in the sentence than those that follow. However, in languages which have articles this difference fades out as we increase the size of the corpus and the estimation of the entropies becomes stable. On the other hand, this is not the case for languages that do not possess articles (Finnish, Estonian, Polish, and Slovak), where the difference persists even after the entropies converge. Thirdly, languages which do not have articles have higher entropy rates than languages which do. While interesting, we cannot attribute this divergence in the presence of articles; the four languages that do not use articles in this study, also have richer case systems which can increase the overall entropy (Bentz, Verkerk, Kiela, Hill & Buttery, 2015).[7] Taken together, the results from the entropy rates show

[6]Formally, bigram language models have high *perplexity*. That is, how well does the language model predicts the sample from the corpus. However, perplexity is equal to $2^{H(T)}$, where $H(T)$ is the entropy of the sequence.

[7]German has a rich case system; however, due to syncretism, only two cases are distinctive.

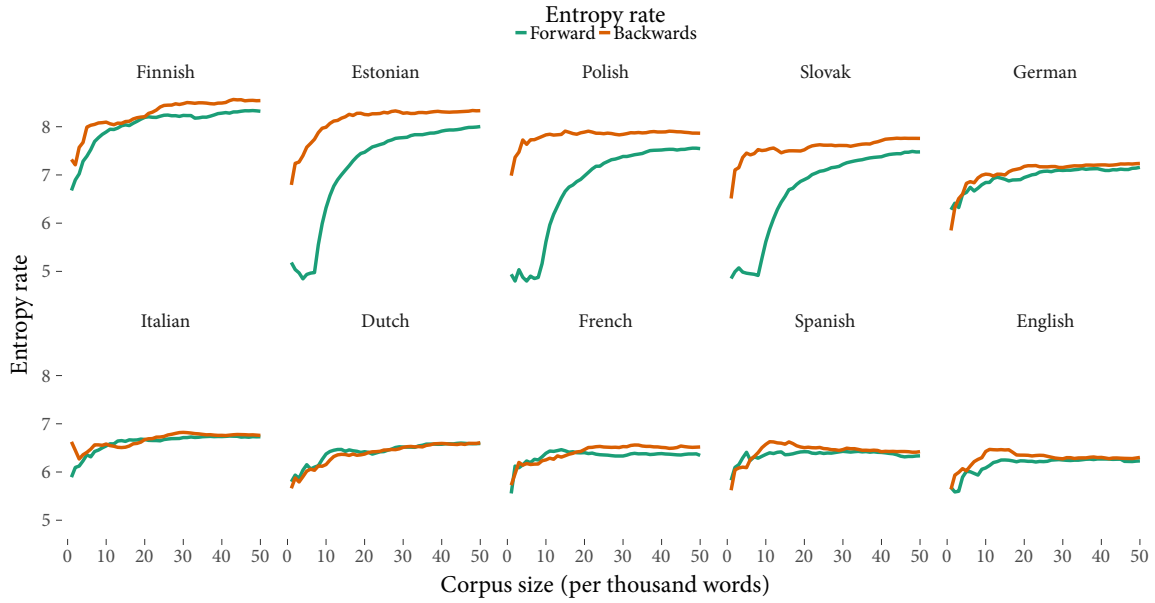


Figure 6.2 Forward and backward entropy rates per thousand words on ten languages from the EUROPARL corpus. Languages are ordered by average entropy rate (aggregating forward and backwards). We compute the entropy rates using the algorithm outlined in Bentz et al. (2017a) on increasingly large sequences for each language stopping at 50000 tokens where all the estimations seem to stabilise. Languages that do not use articles (Finnish, Estonian, Polish, and Slovak) show higher differences between the forward and backwards probabilities compared to the rest of the languages. Possessing grammatical gender, on the other hand, does not seem to relate to those differences (see, Table 6.3, and text for quantitative analyses).

that grammatical constructions might affect the average information content carried by some languages when we look at the words that follow compared to the words that preceded.

Turning to the transitional probabilities, we ask whether more information is carried across languages in constructions of the form *any* ↔ *noun*.[8] To do so, we examine the following independent variables; (1) the name of the language (`language`) which we use as a random effect in the mixed effects model, (2) type of bigram (`bigram.type`) *any* ↔ *noun* and *any* ↔ *any*, (3) whether the language in question uses grammatical gender (`gender`), (4) whether the language uses articles (`article`), and (5) the type of transitional probability (`tp`). We select the variables that enter the model via the *lasso* procedure outlined in §3.3.1. The independent variables which remained were `bigram.type`, `article` and `tp`. We then

[8]We use a forward arrow to note prediction, a backwards for retrodiction, and a bi-directional to include both in cases where the directionality of the relationship is irrelevant.

proceed to construct two models; a simple linear regression model of the form,

$$TP = \beta_0 + \beta_1 \texttt{tp} \times \beta_2 \texttt{bigram.type} \times \beta_3 \texttt{article} + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

where the transitional probability is predicted by the intercept β_0 plus the sum of the products of the three independent variables with their respective slopes plus a normal error ϵ ($\times$ is used to denote the main effect of both variables *plus* their interaction). Since languages differ in their entropy (Table 6.3 and Fig. 6.2), we also construct a similar *linear mixed effects model* with the addition of a random intercept for language.

$$TP = \beta_0 + \beta_1 \texttt{tp} \times \beta_2 \texttt{bigram.type} \times \beta_3 \texttt{article} + \gamma_1 \texttt{language} + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Both models give similar estimates for the coefficients of the fixed effects although the estimations of the linear mixed effects model yield smaller standard errors in general. However, the addition of `language` as a random effect yields higher fit to the data ($\Delta R^2 = 3.4$). We assess whether the added complexity of the model from the addition of a random effect is necessary by comparing the AIC (Akaike, 1974) of each model which ensures the best tradeoff between the number of parameters and the fit to the data. The simple linear model was better ($\text{AIC}_{\text{OLS}} - \text{AIC}_{\text{LMM}} = 29.9$) so we focus on that model for the rest of the results section.

Table 6.4 presents a summary of the main effects and interactions in the ordinary least squares model. The β values show the comparison between one level of the variable and the variable mean. For example, `tp (btp)`$= -0.545$ means that the average surprisal in the backwards transitional probabilities is 0.545 bits lower than the variable mean. Firstly, there is a main effect of the direction of the transitional probability rendering the backwards transitional probability significantly more predictable than the forward (lower surprisal). Similarly, we see a main effect of `type` showing that bigrams where the second element is a noun are more predictable than any random bigram. Interestingly, there was also a main effect of `article` meaning that overall languages that do not have articles are more predictable than languages which do. The picture is further complicated when one considers that the languages without article are the ones with the higher entropy rates (cf. Fig. 6.2). If anything, the more tokens should make these languages less predictable as the entropy is increased (Table 6.3). For example, if we are sampling words at random from a language 'urn', we would be more uncertain about the word we are sampling from the *Finnish* 'urn' than that of *English*. This effect, however, can be interpreted when we consider how the transitional probabilities are calculated. Considering (5.2) and (5.3), English by having fewer word types should have less

unique bigrams as the same higher frequency words keep repeating, giving rise to lower TPs. Conversely, *Finnish* has more unique bigrams composed of less frequent words resulting in higher average TPs.

Fig. 6.3 illustrates the various interactions found in Table 6.4. Figure 6.3a shows the interaction between type of bigram and directionality of the transitional probability. As expected, in the random case, transitional probabilities were almost the same with a slight numerical advantage of FTP, a picture which is, however, reversed in the case where the second element of the bigram is a NOUN, meaning that nouns are more constrained in the contexts in which they can appear. Moving on to Fig. 6.3b we see that languages which have articles have a larger difference between FTP and BTP than languages that do not have articles. We also note here that in this figure we are collapsing over type, therefore, the lower scores are driven partly from that. Arguably, the most interesting interaction is the one between article, type and tp (Fig. 6.3d); forward probabilities are slightly more predictable when looking at any random bigram in the corpus. However, this seems to be modulated both by the type of bigram and by the direction of the transition (tp). Backwards probabilities are significantly more predictable for bigrams where the second element is a NOUN than for any random bigram. While this holds for all languages examined, it seems to be further modulated by languages which use articles where a larger effect is observed.

6.3.4 Discussion

Relating these results to semantic implicit learning, Williams (2005, p. 295) found that participants who spoke languages with grammatical gender, performed better in the tasks. However, except for a few participants who spoke Slavic languages as an L1 most participants spoke *French*, *German*, *Spanish* or *Greek* all of which use articles. Based on the above results we can conjecture that it might have been the fact that participants knew languages which have articles instead of grammatical gender that drove the better performance. The results here add to the bulk of evidence that participants might employ knowledge from their L1 during these tasks (Onnis & Thiessen, 2013). A wider empirical study involving a richer sample of L1s could potentially uncover the relation between BTP and performance in such tasks. However, we cannot make any claims about how experiencing different L1s might shape learning of a gendered L2. If it holds true that learners of gendered languages attempt in a way to minimise some backwards surprisal cost, BTP might emerge as a more objective, quantitative way of predicting the learning curves.

During the feature selection procedure, the independent variable of gender was eliminated. We also examined whether a model which has gender as an independent variable instead of article would be able to provide a similar fit to the data. The fit is considerably worse

Table 6.4 Summaries of the OLS and the LMM predicting the surprisal values. The values displayed are the β coefficients (one standard error above or below the mean in parentheses). The β values show the comparison between one level of the variable and the variable mean. The level in parenthesis corresponds to the β value and the levels are the same for the interactions. The two models were trained with on same predictors with the exception that in the linear mixed effects model we added a random intercept for each language.

	OLS	LMM
Constant	6.582^{***}	6.582^{***}
	(0.0677)	(0.0856)
tp (btp)	-0.5451^{***}	-0.5451^{***}
	(0.0677)	(0.0605)
type (nouns)	-0.4809^{***}	-0.4809^{***}
	(0.0677)	(0.0605)
article (no article)	-0.6633^{***}	-0.6633^{***}
	(0.0677)	(0.0856)
tp × type	-0.743^{***}	-0.743^{***}
	(0.0677)	(0.0605)
tp × article	0.2566^{***}	0.2566^{***}
	(0.0677)	(0.0605)
type × article	0.0622	0.0622
	(0.0677)	(0.0605)
tp × type × article	0.3741^{***}	0.3741^{***}
	(0.0677)	(0.0605)
Observations	40	40
R^2	0.931	
Adjusted R^2	0.916	0.95[1]
Log Likelihood	−17.52	−31.463
Akaike Inf. Crit.	53.05	82.92
Bayesian Inf. Crit.	68.24	97.58
Residual Std. Error	0.42 (df = 32)	
F Statistic	61.49^{***} (df = 7; 32)	

Significance levels: $^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Note: tp = Transitional probability (forward, backwards); type = Bigram type (random, noun in second position); article = Whether the language in question uses articles.

[1] The coefficient of determination for the linear mixed effects model was calculated using the method outlined in Jaeger, Edwards, Das & Sen (2016).

regarding the fit to the data ($\Delta R^2 = .29, B_{ag} = 1.96 \times 10^8$) as well as in AIC values (53.05 for the model with the article covariate vs. 113.15 for the model with the gender convariate).

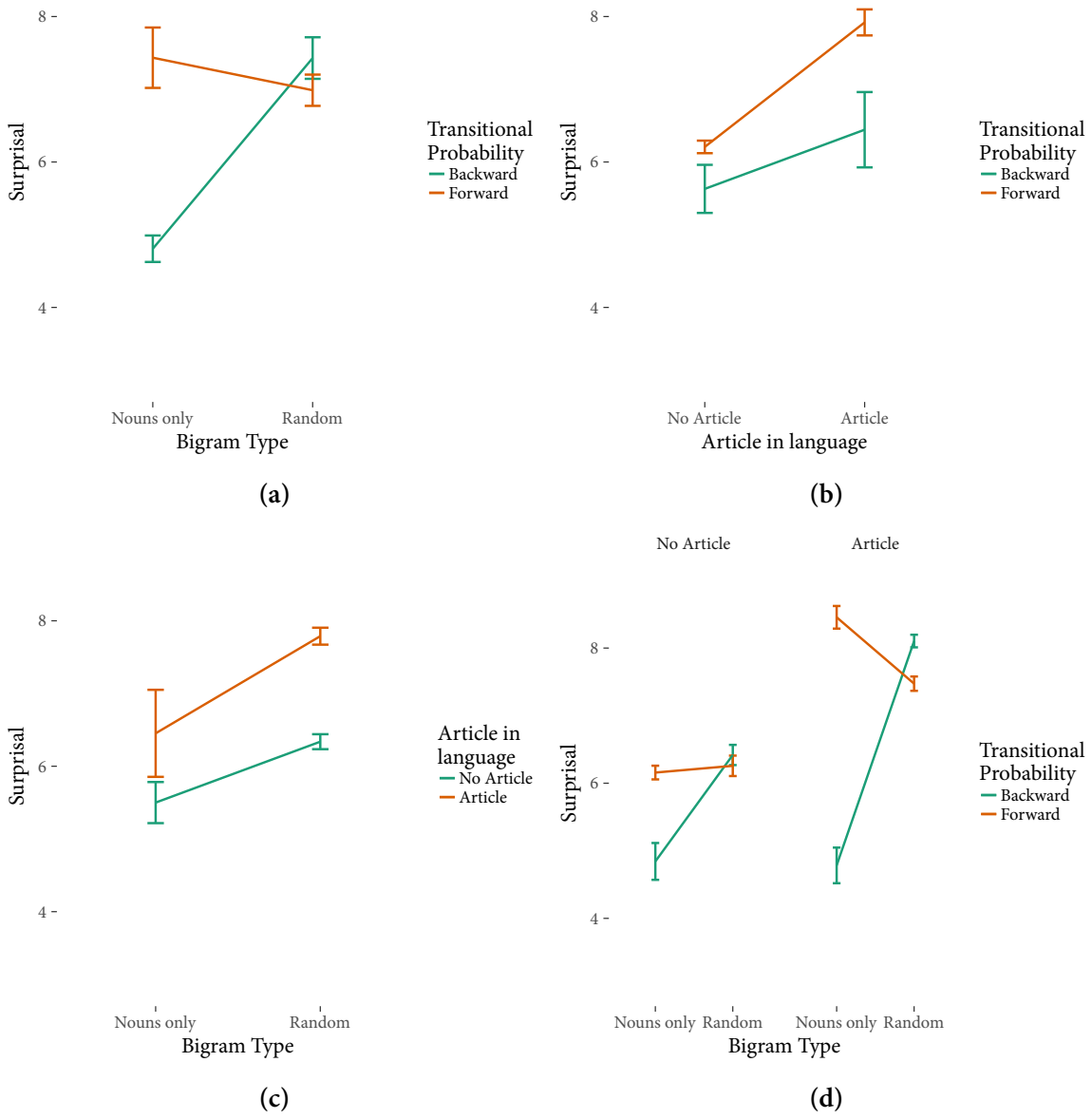


Figure 6.3 Three way interaction between `tp` × `type` × `article`. Forward probabilities are slightly more predictable in general (see text), however, this seems to be modulated both by the `type` of bigram and by the direction of the transition. Backward probabililities are significantly more predictable for bigrams where the second element is a NOUN than for any random bigram. While this holds for all languages examined, it seems to be further modulated by languages which use articles where a larger effect is observed.

These results point to the direction that the retrodictive cues inherent in a low-level statistical regularity, such as having an article before a noun, might be responsible for biasing the way learners perceive ARTICLE ↔ NOUN groupings. The fact that grammatical gender (or its lack thereof) by itself cannot explain these patterns better than the presence of articles is interesting as it points to the direction that speakers of a language do not need to have access to grammatical information over and above surface statistical cues. Relating this finding to the above discussion, we predict that speakers of languages such as Polish or Slovak which mark grammatical gender by morphological suffixes (e.g., feminine nouns end in *-a* in Polish) would perform worse in the generalisation task than speakers of languages with articles.

In the present study, we focus on the statisical bias language speakers may have because of the higher backwards transitional probabilities in the cases where the language has articles. However, this does not mean that only ARTICLE → NOUN relations can introduce such biases. Onnis & Thiessen (2013) through corpus analyses show that in the case of PREPOSITION → NOUN bigrams backwards probabilities are also more important than forward for reasons similar to the articles. However, while languages that do not have articles are quite common, languages without some form of adposition are less common. Indeed, in the World Atlas of Language Structures (WALS) dataset (Dryer & Haspelmath, 2013) from the 2679 languages 1153 use some form of adposition (preposition, postposition, inposition, or without dominant order), whereas only 30 do not use adpositions.

The measures implemented here are influenced by the rich tradition in the statistical learning literature (e.g., Saffran et al., 1996) which holds that the information held in the transitional probabilities is important in segmenting words from the input (also Swingley, 1999 and Goldwater et al., 2009). These measures have been found to correlate with reading and processing times (Frank et al., 2013), to be predictors of cloze probability tasks (Shaoul, Baayen & Westbury, 2014) as well as dominating many practical NLP applications (e.g., Ganesan, Zhai & Viegas, 2012). While backwards transitional probabilities are not as widely used a measure (Onnis & Thiessen, 2013; Pelucchi et al., 2009; Swingley, 1999, for applications), we have already noted that current bi-directional recurrent neural networks (Schuster & Paliwal, 1997) used in speech recognition (Graves et al., 2013), text reading (Hermann, Kociský, Grefenstette, Espeholt, Kay, Suleyman & Blunsom, 2015), sentiment analysis (Tang, Qin, Feng & Liu, 2016) use both a forward and backwards pass for each sentence, effectively learning both the forward and the backwards probabilities. Using this mechanism, these RNNs consistently achieve higher scores in many tasks than their simpler uni-directional counterparts. A simple explanation for this would be that the model has seen more data hence its performance is improved. However, the above experiments show that while this might be the case, information contained in the backwards pass is also qualitatively different in the case of bi-directional RNNs.

Chapter 7

General Discussion

In this final section, we summarise the main findings of the present thesis in a way to support our initial hypothesis that semantic implicit learning is a necessary by-product of language processing. Furthermore, we offer one final study that mostly aims to bridge a gap between the current results and more general questions explored by cognitive scientists. Concretely, we have argued so far that the distributional statistical patterns inherent in the linguistic structures bias learners in a way that makes some constructions easier or harder to learn. However, while significant, this gives a small advantage to language learners. In our last study, we explore whether knowledge of the distributional regularities can translate to 'deep' semantic knowledge. We conclude by outlining the implications of this research and some directions for future study.

7.1 Summary of findings

Chapter 3

Studies 1 and 2 sought to find a principled way of deriving appropriate semantic representations for modelling tasks of semantic implicit learning. We did so by comparing the performance of several different models of semantics on tasks of semantic priming, which, as we argued, bear many similarities with SIL tasks. Through rigorous analyses of the performance of the different models on a large dataset of reaction times and studies of mediated semantic priming, we find that a particular class of models, the neural embeddings, outperform the rest.

Chapter 4

Studies 3 and 4 extend the results obtained in the previous chapter by measuring the effect of semantic neighbourhood density, as computed by the models in the earlier studies, on lexical decision reaction times. We argue that focusing on this metric yields two advantages; firstly, we can look past pairwise relations (as in the case of semantic priming tasks) to the global configuration of concepts in the semantic space. Secondly, because datasets of lexical decision reaction times exist in languages other than English, we are able to explore whether the psychological validity of these measures extends there as well. We find that the neural embeddings still outperform the other models and that they also improve the fit of regression models predicting reaction times.

Chapter 5

Studies 5–8 apply the semantic representations on tasks of semantic implicit learning. We start by outlining a computational description of the SIL tasks and the appropriate modelling framework for capturing the effects observed during the behavioural experiments. Subsequently, we look at four different semantic distinctions; two 'core' ones *animacy* and *concreteness* and two perceptual ones. We also compare the predictions of the models on speakers of two different L1s; English and Chinese. We theorise that the behavioural patterns are consistent with a distributional approach to semantics and that non-distributional models of semantics would find it difficult explaining the generalisation gradients of the human participants. The simulations offered there support these ideas.

Chapter 6

Studies 9 and 10 sought to verify two assumptions we had made throughout the present thesis. Firstly, that surface-level phonological information cannot explain the results of the experiments, and, secondly, that *backwards* probabilities can be just as informative for the learner, but also qualitatively different. In two studies we find that phonological information cannot explain the patterns observed in the data but might be responsible for the low generalisation scores usually seen in these experiments. Secondly, we find that across languages, backwards probabilities can be quite informative when we consider specific grammatical relations. These results might help predict individual differences in SIL tasks.

7.2 Study 11: Beyond co-occurences?

7.2.1 Introduction

In §2.3.1 we outline a frequent criticism against models of distributional semantics which usually centre around the question of how these representations can relate to our knowledge of semantics. In other words, while it is interesting that the words *cat* and *fur* are somehow related distributionally, there is nothing in the semantic representation of the word *cat* that indicates that 'cats have fur', instead of 'cats eat fur'. Although some analogical reasoning of the form `man:king::woman:?`[1] is possible in this class of models (Mikolov et al., 2013d), these models indeed lack any knowledge of semantic relations present in other semantic representations (see §1.3.1). Nevertheless, we have seen throughout the current thesis that such distributional representations can capture effects observed in behavioural data which other models find difficult to explain.

Even though using such representations has proven quite effective in modelling semantic priming phenomena as well as patterns of semantic implicit learning, the scope of distributional semantics still seems quite limited from a cognitive point of view. That is to say, for a critic, all this provides is, at most, a proxy of the *meaning* of a word which we can use as a computational device to simulate empirical phenomena. However, this would still be unrelated to cognition, and how we acquire semantics, semantic relations or how it is placed along psychological theories of meaning. In this section, drawing on previous work within the *parallel distributed processing* framework (McClelland, McNaughton & O'Reilly, 1995; Rogers & McClelland, 2004; Rumelhart & Todd, 1993), we consider extensions of such models using distributional representations. Our aim is to show that instead of simple devices, distributional models of meaning can serve as building blocks, *bootstrapping* a semantic system providing *prior* knowledge to a model aiming at capturing how semantics are learnt.

Rumelhart & Todd (1993) explore the derivation of distributed word representations (i.e., real-valued vectors) from associating one-hot representations with semantic properties. Figure 7.1 shows the model used in their simulations. Inputs to the network were only a *localist* representation along with an also *localist* relation representation. These two inputs predicted a semantic property in the output layer. For example, if the unit of *cat* was activated with the unit signalling the IS-A semantic relation the node for *animal* should activate. By back-propagating the errors to the representation layer, the network ends up learning word representations which are predictive of the semantic properties modulated by the semantic relations. McClelland et al. (1995) and Rogers & McClelland (2004) extended this work by providing a fully-fledged theory of acquiring semantics by showing how such networks can

[1] These analogical formulae should be read as: man is to king as woman is to ....

be used to capture qualitative empirical patterns such as basic level preferences, dementia, and frequency effects.

An immediate problem with these models is their apparent inability to generalise to novel words. While the patterns observed during training might follow the behavioural patterns found in the literature, generalising to novel words means that we need to re-train the model to capture the predictions for the new words. Here, we examine whether distributional models of semantics can serve as a starting point for such an associationist model. Having already a consistent way of deriving word representations, associating them with semantic properties would enable us to check whether the model would be able to generalise those properties to other words that have not entered the system yet. If such a hypothesis turns out to be true, it means that the distributional vectors capture something more than a compressed version of the word's linguistic environment. From a computational cognitive point of view, they could provide a *prior* distribution of weights we need to set to learn semantics.

7.2.2 Materials and Methods

Data

We use the extended training corpus from Rogers & McClelland (2004, Chapter 3 and pp. 396–397) which provides a more complex set than that used originally by Rumelhart & Todd (1993). This training set contains 21 nouns, either plants or animals, and four contextual relations; IS-A, IS ..., CAN ..., and HAS .... The properties chosen for each noun are indicative of those available to infants either through direct perception of through verbal statements. As elsewhere in the current thesis, we use the neural embeddings from §3.4 as our word representations. For each relation and property we generate localist representations which we then use together with the neural embeddings as input/output patterns to the network.

Model

For the following simulations we adopt the architecture used by Rumelhart & Todd (1993) extending the architecture illustrated in Fig.5.1 by adding one more input layer feeding to the *hidden* layer independently of the *representation* layer. This layer feeds semantic relations to the *hidden* layer such as *has, is, can*, essentially modulating the meaning representation before making any predictions about the semantic properties. Algebraically, this modulation is tantamount to shifting the raw word representations in such a way that given the same word input but different relational input, the output layer will have different activation patterns. Regarding the neural embeddings, we follow the method outlined in §5.3.1. More specifically, we use *localist* representations in the input layer, but we substitute the weights of the input →

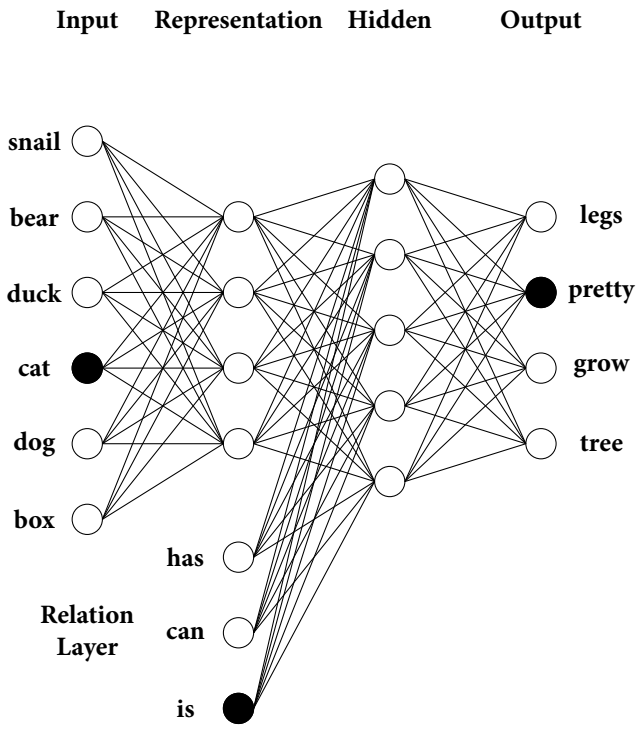


Figure 7.1 The architecture of the network used by Rumelhart & Todd (1993) and Rogers & McClelland (2004) to associate semantic relations (e.g., IS-A) with semantic properties (e.g., *grow*). As a computational graph, this network is an extension of the one used in §5.3.1 to predict performance in implicit learning tasks with the addition of a *relation* layer. For our baseline (random) model we use the same initialisation procedure outlined in Rogers & McClelland (2004). For the pre-trained vectors, the size of the representation layer is the same as the size of the distributional vectors (i.e., 300 units). Due to space constraints we cannot depict the entire network, however, for the pre-trained vectors, the size of the input layer was the size of the vocabulary $|V|$, the size of the representation layer is 300, the hidden unit size is 50 and four relation units.

representation matrix with those of the neural embeddings. The dot product of the localist input with the input → representation weights will effectively activate the distributional representation of each word in the representation layer (note that we do not use non-linearity in this layer).

7.2.3 Results and discussion

Fig. 7.2a shows the by-epoch total cross-entropy error. Using the pre-trained distributional representations as input to the model the computational time needed to minimise the training

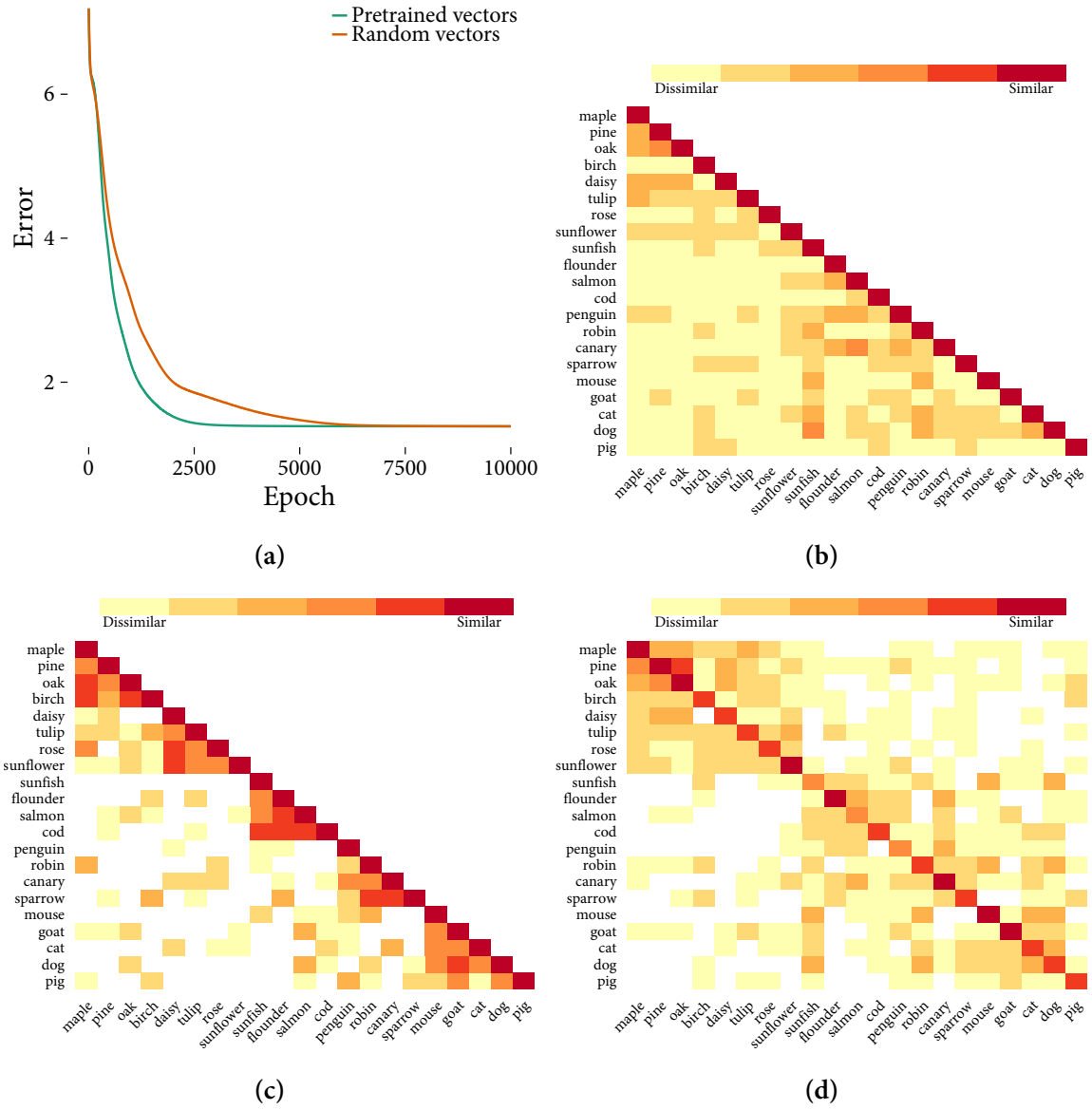


Figure 7.2 (a) Cross-entropy error curve on the training data as a function of epoch. The pre-trained distributional vectors (green) show an advantage as they level considerably faster than their random counterparts (orange). (b) Similarity matrix for the distributional vectors. While some clusterings are visible (plants vs. animals) there are no basic level effects. (c) Similarity matrix for the trained distributional vectors (i.e., the vectors used before but having been associated with their semantic properties). While the more general clusterings are retained, we see a clear basic level effect. (d) Similarity matrix between the old (distributional) and new (trained distributional) vectors. The diagonal of the matrix shows how much each word representation changed from the training procedure (i.e., how similar is the purely distributional to the trained distributional vector).

error is reduced considerably from 6000 (random initialisation) epochs to 2500 (distributional vectors). This speed-up can be considered as giving the model a better starting point (i.e., *bootstrapping* the system) in the sense that it already knows something about the words used in training. However, we need to be cautious here as the comparison is not equal because the size of representation layer differs between the two models. When using the distributional vectors, the size of the representation layer becomes the size of the vectors (i.e., 300 elements). On the other hand, using the random vectors, we follow Rumelhart & Todd (1993) and set the size of the representation layer to 15. While, therefore, the number of parameters in the model increases considerably in the first case from 2555 (random word-vectors) to 8540 (pre-trained vectors), it might be the case that the network is trying to compress the output in denser representations. The apparent *bootstrapping* can also be related to the idea of *curriculum* learning, which has recently been started to be explored in this context (Tsvetkov et al., 2016). According to this, the network benefits from a better initialisation of weights as it brings it closer to its final solution than a more random initialisation.

Fig. 7.2b shows the pairwise similarity estimates between the neural embeddings in the training corpus. While we can see some clusterings emerging, as there is greater within category similarity for both plants and animals than between category, these relations become clearer when we look at the same representations *after* training them to predict the semantic properties. In Fig. 7.2c we plot the pairwise similarities for the same embeddings extracting the input → representation matrix *after* the training of the network. We see that the distinctions become more prominent in this case as the types of trees, flowers, fish and other animals cluster more closely than before. Interestingly, while for the trees, flowers, and fish there were specific properties in the output layer that marked them as such, in the case of animals, there were none. The greater within-cluster similarity for those nouns has presumably been inferred from common properties of those nouns such as that they have legs, fur or that they can walk. Finally, Fig. 7.2d shows the similarity between the semantic representations before and after training the network to predict the semantic properties. Our main interest is on the diagonal of this matrix which shows the distance for the same noun before and after training. We see that the resulting embeddings do not differ a lot from the original ones indicating that the network only highlights regions of interest in the input minimising the effect of irrelevant 'noisy' neuronal ensembles.

Assessing the generalisation patterns

Initialising the representation weights randomly has the shortcoming that we cannot see how would the model fare on unseen instances. While this is important in understanding how we form semantic categories, how we learn different semantic properties features, or how we

generalise to other features, it falls short in predicting the semantic properties of unseen words. Under this paradigm, any to-be-learned word has to pass through a similar procedure where the network needs to learn the semantic properties through extensive exposure. However, knowing that *rose* IS-A *flower* and that *daffodil* is very similar to *rose* in some sense, we might be inclined to believe that the properties of *rose* extend to *daffodils* as well.

Generalising in this manner would uncover a compelling role for distributional representations as it would take much of the computational cost needed to learn semantic properties from scratch every time a word enters the mental lexicon. *Bootstrapping*, therefore, would extend to unseen words as the already trained words would pave the 'path' to learn the properties of the novel ones. In the previous section, the network learned to associate features of the input vector (the representation layer) with semantic properties as patterns of activation in the output layer, modulated by the semantic relation representation. If in similar regions of the input layer it can find the features that became predictive of the properties, it should be extendable to unseen —yet similar— words.

We noted above that the input representations (the distributional vectors) remained as free parameters in the network meaning that during training they could be updated to accommodate the new input. Since the optimisation method was a gradient descent variant, the input representations were updated to accommodate for the loss in predicting the correct output properties. Concretely, the representation matrix $\mathbf{M}_{\theta_w}$ has been iteratively updated as $\mathbf{M}_{\theta_w} = \mathbf{M}_{\theta_w} - \eta \cdot \nabla J(\theta_w)$.[2] This procedure means that while the raw word embeddings are in the same vector space as their trained counterparts, we need to *translate* the raw embeddings into *semantic representations*. *Translation* is an affine transformation whereby we move each point in the vector space in a given direction. One way we can transform the raw representations is by keeping track of the linear maps (the gradient of the error function w.r.t the word matrix) applied at each update step to the word embeddings and then transform the raw distributional representations to semantic ones. However, this procedure was difficult to achieve in this context as we update the weights of the network in batches instead of after the presentation of every word.

We achieve generalisation to novel instances by approximating the translation operation by learning a function $F : W \rightarrow S$ which maps word embeddings as constructed by any DSM to a *semantic representation* which contains information about the different relations and properties this word enters. Under this function, mapping a distributional word vector to one that contains semantic information relations is a linear operation $f(\mathbf{s}_i) \approx \mathbf{M} \cdot \mathbf{w}_i + \mathbf{b}$ where $\mathbf{M}$ and $\mathbf{b}$ are the free parameters needed to be learnt, $\mathbf{s}_i$ the learnt semantic representation,

[2] This is the general formulation of gradient descent based algorithms, and we use this here for simplification. As noted above, the algorithm used was ADADELTA (Zeiler, 2012) which adaptively changes the learning rate η.

and $\mathbf{w}_i$ the word embedding. Following this procedure, we can 'trick' the network by feeding representations directly to the representation layer by applying nonlinearity to the output of the above function $\sigma\big(f(\mathbf{s}_i) + \mathbf{b}\big)$.

We evaluate the performance of the model by keeping track of the activations in the output layer once a (word, relation) tuple enters the network. Typically, we round activation values over 0.5 or 0.75 to 1 (i.e., the neuron activates) otherwise 0 (i.e., the neuron does not activate). A problem which we encountered was that for the 're-constructed' inputs more neurons activated than it should. We attribute this behaviour to the fact that even minute differences can become drastic for the network. We correct this inconsistency using a thresholding parameter τ which we increase until we find the activation that a neuron needs to have to reproduce the training patterns. In other words, we feed the raw training vectors to the function, and we increase τ until the activation patterns match the teacher patterns. We find that a value of $\tau = 0.9$ worked best and we report the simulations based on this value.

Table 7.1 shows the predictions of the model on eight novel words from the same categories as those used by Rogers & McClelland (2004). The model can accurately capture the more general properties such as IS-A *plant* or *animal*, occasionally returning correct predictions for the other properties. One interpretation of the results in this table is that the network has failed to predict the more specific properties. However, we note here that the high thresholding parameter might mask some true positives. Indeed, lowering the threshold increased the number of properties predicted at the cost of also predicting unrelated properties. We chose, therefore, to continue with the high threshold. One interesting prediction of the model was that it classifies the word *rat* as a *mouse* in the basic-level category. This behaviour might lead to an erroneous analogy that `elm:tree::rat:mouse` (i.e., that *elm* is to *tree* as *rat* is to *mouse*). However, there is a simple explanation to the above; following the classic findings of Rosch, Mervis, Gray, Johnson & Boyes-Braem (1976) on the different levels of specificity Rogers & McClelland (2004) associated *mouse* with *mouse* as a basic level property and *animal* as superordinate. Trees, on the other hand, as was found by Rosch et al. (1976) were considered as a basic level category by North American subjects, so *elm* was one level lower than *tree*. In the present context, since *rat* and *mouse* are very similar concepts, the model makes the same inference for *rat* as it does for *mouse*.

A problem in this short of qualitative analysis is that the examples taken may run the criticism that they are 'cherry-picked' (i.e., we selected the only 12 examples which agree with our predictions). This objection might be countered by the fact that the training set was insufficient for the model (four *plant* words and six *animal*). However, it would be interesting to see how it would fare on a larger scale. We examine this by focusing on IS-A relations which we can trivially extract from WordNet. Table 7.2 shows the semantic categories used and the

Table 7.1 Model predictions for the semantic properties of a set of novel words from the same basic level categories as the ones used during training.

Semantic Category	Noun	Relation					
		IS-A_G	IS-A_B	IS-A_S	IS	CAN	HAS
TREES	ELM	*plant*	*tree*	*plant*			
	FIR	*plant*					*roots*
FLOWERS	DAFFODIL	*plant*				*grow*	
	POPPY	*plant*	*flower*			*grow*	
ANIMAL	RAT	*animal*	*mouse*		*living*		
	TORTOISE	*animal*			*living*		*skin*
BIRD	PIGEON	*animal*			*living*		*skin*
	CHICKEN	*animal*			*living*		
FISH	MACKEREL	*animal*					
	TROUT	*animal*					

Note: G = general; B = basic; S = specific. Empty cells indicate that the network did not activate any property nodes for that relation.

corresponding WordNet synsets. We evaluate the performance of the model by keeping track of the activation of the corresponding property in the output layer. If the model activates *only* the correct property for a specific relation, then the prediction is correct, otherwise wrong. For example, given *elm* and IS-A_G, activating only the *plant* property would be correct but activating *plant* and *tree* would be incorrect. Table 7.2 shows the accuracy rates for each category by semantic relation. The scores for more general categories are higher indicating that they are more well-represented in the distributional vectors. While the results from the more specific categories appear considerably lower, we note that the model had seen only four positive examples of each subcategory.

These results can have a simpler explanation; the model has a bias towards the *animal* property, therefore, continuously activating this node would lead to high accuracy rates. Such a bias could be either the result of how the distributional vector represents properties related to animals or reflecting the fact that there was an asymmetry in the training data which contained more animals than plants. We test this by repeating a similar procedure selecting from WordNet 700 *animal* words, 700 *plant* words and 700 words not falling into either category and repeat the experiment. For the *animal* and *plant* words, the model should activate the corresponding node in the output layer, whereas for the random baseline it should not activate anything (a correct response is then considered if no node is activated). This way

Table 7.2 Proportion of correct activations of the corresponding semantic property node in each category.

	Semantic category	WordNet synset	# elements	Mean[1]
IS-A$_{\text{general}}$	*animal*	`animal.n.01`	1389	0.7 (0.01)
	plant	`plant.n.02`	758	0.39 (0.02)
IS-A$_{\text{basic}}$	*flower*	`flower.n.01`	24	0.21 (0.08)
	tree	`tree.n.01`	69	0.2 (0.05)
	fish	`fish.n.01`	99	0.11 (0.03)
	bird	`bird.n.01`	181	0.33 (0.03)

Note: Ω in this context is the set of all WordNet synsets.
[1] The numbers in parentheses denote standard deviation of the mean. Since the evaluation was done on hits and misses, standard deviation in this context is calculated as the binomial standard deviation $\sqrt{\frac{p(1-p)}{n}}$.

we ensure not only that the model activates the correct nodes when it should but also that it correctly rejects words it cannot classify.

Table 7.3 shows the scores of the model on the *precision*, *recall* and F_1. In short, *precision* measures the proportion of correct responses against those returned, whereas *recall* measures the proportion of correct responses against those that should have been returned. The two recall scores correspond to the scores in Table 7.2 showing that the samples taken were representative of the above estimates. Interestingly, the *precision* is quite low for the *animal* category and higher for the *plant*. This results indicates that the model will be more biased to select *plants* than *animals* but when it does select animals chances are it will be correct. As for the random condition, the scores are considerably lower, meaning that the model incorrectly activates either of the two general categories.

While this behaviour seems problematic, we note that the model has been trained to predict only two concrete categories; considering that the size of these categories is quite limited compared to the entire WordNet vocabulary (75221 synsets), we can conjecture that during training the model would consider more features of the input as relevant. If we had included another IS-A$_{\text{G}}$ *abstract* property in the model, or restricted the extension of the random set to include only concrete objects, then the model predictions might have been better. Since humans acquire concrete concepts easier than abstract ones (Paivio, 1971), this behaviour poses a unique challenge to this kind of models which can be examined further in future research.

Table 7.3 Precision, Recall and F_1 scores on two semantic categories and random baseline.

Semantic Category	WordNet synset	Precision	Recall	F_1 score
plant	`plant.n.02`	0.74	0.39	0.51
animal	`animal.n.01`	0.56	0.69	0.62
random	$\Omega \setminus (\texttt{plant.n.02} \cap \texttt{animal.n.01})$	0.36	0.45	0.40

Visual inspection of what the model predicts for the random condition might indicate the solution it found during training. To see this qualitatively, we probed the model with words that appear in the semantic neighbourhood of the trained words but did not possess any of the above properties. For example, *keyboard* is related to *mouse* as a computer device, or *sausage* as related to *chicken* as they are both edible. Crucially, none of the words have any properties which can be found in the output layer. As conjectured above, the model activates the *animal* property for the IS-A_G relation (via *chicken*) or in the case of *keyboard* the *mouse* for the IS-A_B relation. This behaviour can be taken to be supportive of the above hypothesis as in the cases where the model does not 'know' what to answer, it has a distributional bias.

7.2.4 General Discussion

The above study sought to find a way to translate 'shallow' semantic knowledge which is gathered by attending to the distributional patterns of words in the language to 'deeper' semantic representations which encode semantic relations. We do so by extending a current model of semantic knowledge which given initially random representations for words and some salient properties, ends up learning meaning representations that encode these properties when modulated by the corresponding relation. Our extension included firstly feeding the network distributional vectors instead of random ones providing a better starting point during training. While our results match those of the existing literature, we further extend them by learning a function which maps raw distributional vectors to 'deep' semantic representations. Following this procedure enables us to correctly classify several novel instances only from the presentation of a handful of training items.

One possible source of criticism is the addition of the function f that maps a 'raw' word embedding $\mathbf{w}$ to the learnt representation $\mathbf{s}$ carrying the knowledge about the semantic relations. However, how do learners acquire f if not by more training? In our simulation, this happens separately *after* the network has been trained to make the relational associations. An immediate problem with this approach is that it seems unlikely that the learners after learning

the relations compare their internal distributional representation to **s**. This would imply that learners can access the contents of **w**, which we argue above is not possible as the distributional representations are causally efficacious *without* being amenable to consciousness.

We see two possible solutions to the above problem; firstly, learning the parameters **M** and **b** of the function is tantamount to adding one more linear layer to the network in Fig. 7.1. In this argument, the weights of the word embedding matrix remain invariant during training and what we update is a trainable copy. This procedure would expand the parameter space of the network rendering it hard to generalise from only a few examples. An elegant alternative is reminiscent of sleep consolidation and is given by algorithms like the 'wake-sleep' (Hinton, Dayan, Frey & Neal, 1995) which asserts that when there is no bottom-up input, the network will attempt its top-down reconstruction minimising the noise in the weights. The resulting weight changes will also affect the untrained word embeddings as the already learnt knowledge will extend to them as well.

The results presented in this study are indicative that distributional semantic representations can serve as the building blocks of learning semantic relations. Xu & Tenenbaum (2007) argue that one of the shortcomings of the associationist models in word learning is that they need a significant amount of data and extensive training to achieve adequate generalisation accuracy as they lack a way of detecting *suspicious coincidences* (see also Griffiths & Tenenbaum, 2007). For example, in the original Rumelhart & Todd (1993) model, for every new word we encounter, inserting it into the network would mean that we would need to re-train its weights to predict its semantic properties. Despite the effort by some researchers, (Regier, 1997, 2003) it seems that such models are unable to overcome the problem of fast generalisation from only a few training exemplars. For a Bayesian model, this would not be an issue as we would exploit some prior knowledge about the word (e.g., 'is a bird') and then extend its meaning to the meaning of the bird. Having, therefore, some *prior* knowledge (e.g., that 'birds can fly') we would rank this hypothesis higher instead of re-learning this association.

One of the problems in Bayesian cognitive models is how do we 'learn' these priors which subsequently guide our inductive inferences (Bowers & Davis, 2012). While many researchers have attempted to answer the question (Griffiths et al., 2012; Hohwy, 2013), the consensus is that the initial contact with the environment somehow shapes our prior knowledge. That is to say, the *input itself* provides the information that subsequently shapes our inferences. During subsequent contact with the environment, these prior beliefs are refined to match current nuances in the input. This constant conflict between our prior expectations and the current input lies at the core of Bayesian models of cognition and is thought to drive many different aspects of human cognition (Tenenbaum, Xu, Perfors & Griffiths, 2011b).

In our simulations, replacing the random weights with pre-trained vectors is tantamount to ascribing some *prior* knowledge to the neural network. During training the network tries to maximise the likelihood $p(\mathcal{D}|\vec{\theta})$, that is, learn the set of parameters ($\vec{\theta}$) that make the data $\mathcal{D}$ more probable. In the random case, these parameters are sampled from a uniform (uninformative) distribution not to bias the network in any way. Conversely, using pre-trained vectors, we bias the network by letting it know some knowledge we already have about the words, namely how they are distributed in language. Models of distributional semantics come in handy here as they can capture exactly that sort of information and alleviate much of the computational cost needed to relearn information that we already know. Relating this to the above discussion regarding the strong and weak distributional hypothesis, we argue that even if the distributional patterns of words in a language do not directly cause the semantic space to be structured in a certain way, they initialise it in a way that is later more amenable to incoming knowledge.

From an engineering point of view, a 'smarter' random initialisation procedure might achieve similar results without the computational cost of pre-training. However, using pre-trained weights to carry out a classification task, has been a quite common practice in machine learning and NLP. In a landmark paper, Hinton & Salakhutdinov (2006) found that pre-training the weights of a neural classifier using a series of autoencoder networks has produced a significant boost in image classification accuracy.

7.3 Implications of current findings and future research

The findings reported in this dissertation have several implications for our understanding of implicit language learning and computational models of cognition. Firstly, the models of distributional semantics introduced in Chapter 2 acquire semantic-like knowledge from the massive amount of linguistic input available to language users. Concretely, the neural embeddings (§2.3.1), the model that consistently outperformed the rest, learn by continuously predicting the context in which each word can appear. Theoretically, these models give an advantage to theories that place learning as a by-product of *processing* (Christiansen & Chater, 2016a; Cleeremans, 2014). If efficient processing is the goal, then predicting the *fitting* context is a means to that goal. A by-product, however, of this prediction is the acquisition of semantic-like knowledge.

Practically, the implication of these results is that semantic implicit learning might not be semantic at all; it is simply learning on the basis of things which, in the mind of learners, are somehow similar. From an application point of view, this poses a unique challenge; how can we transform the input that the learners see in a way such that we maximise the learning gains?

This is another instantiation of the curriculum learning idea introduced above (Tsvetkov et al., 2016). The goal of the teacher now is not to give sufficient examples to the student but present them in the order that would preclude the student from falling into local minima.

Secondly, the four studies seeking to find the most appropriate semantic representations for the SIL tasks can have profound implications for subsequent modelling work. So far, most cognitive modelling work involving distributional semantics has either used the publicly available vectors from Landauer & Dumais (1997)[3] or in-house semantic representations on non-publically available corpora. However, as we have argued in the former case, such vectors might not be appropriate for the present task as they yield low fit to the tasks or in the latter it would be difficult to compare the results of those studies with other datasets. In an attempt to correct this issue, Mandera et al. (2017) have also made their vectors publically available which can also be used in these tasks. In §3.4, we find that our data splitting and validation procedure yields a better fit than other models. Furthermore, the simulations in §4.5 offer similar semantic representations for four languages other than English; Chinese, Dutch, French, and Malay. Such representations can aid subsequent research seeking to explore language modulation on the performance in cognitive tasks or potential L1 biases when learning English. All the semantic representations used in the above experiments are made publically available for subsequent comparisons.

Thirdly, the study offered in §7.2 shows a new potential for research. Most work in cognitive research has dismissed the potential of distributional models of semantics and has opted for either feature or association norms, or similarity rating studies. The reasons for this stance have been briefly explored in various sections of the present thesis but they can be summarised as follows; (1) such models are computationally costly, and they assume mechanisms not implemented in the brain's cortical structures (e.g., dimensionality reduction via SVD), (2) they are not grounded in perceptual experience, and (3) they implicitly assume that speakers of different languages will have distinct semantic representations, an idea that cognitive psychology has long rejected. In what follows, we will examine how the present thesis has addressed these issues.

Firstly, within the context of neural embeddings we see that the reduction in the number of parameters alleviates much of the computational cost associated with earlier models such as LSA or HAL. Secondly, we have stressed in various points that generating semantic representations solely from textual corpora is but one way of learning semantic associations (Kiela et al., 2016). While relevant to the present tasks we can assume that such knowledge is enrichened by other sources of information (Andrews et al., 2014, 2009). Thirdly, §7.2 shows that (3) does not need to be the case; the co-occurrence patterns of words can be

[3] Available at `http://lsa.colorado.edu`

thought of as a *prior* distribution of possible meanings for novel concepts which are then refined by experience. This explanation provides a middle ground between work on cognitive psychology (Heider & Olivier, 1972) which shows that speakers of different languages share semantic representations as these are shaped by their contact with the world and current work exploring minor differences and biases between speakers of different languages (Cibelli et al., 2016, for a recent example). These results are also consistent with computational work on the mechanisms of word learning (Xu & Tenenbaum, 2007). This line of work seeks to understand how speakers can infer the meaning of words from only minimal exposure. Associationist accounts have severe limitations on this issue, as they need an extensive amount of training and data. The solution given in Xu & Tenenbaum (2007) is to assume that prior knowledge biases humans towards certain hypotheses reducing the cost associated with neural network accounts of traversing the entire hypothesis space. In this study, we argued that knowledge of the distributional patterns of words can be thought of as the prior knowledge needed to infer the meanings of new words.

Why do cognitive psychologists persist in reusing the examples of implicit learning outlined at the beginning of Chapter 1? What possible link could there be between judging the trajectory of a ball or learning a natural language? Despite numerous advances in cognitive modelling over the past three decades, we are still quite far from answering these questions. Connectionist modelling attempted to link these phenomena by presenting a 'general-purpose theory of cognition' (cf. Fodor, 1983) by looking at how the brain responds to input: "We do not need to know much if we have simple learning mechanisms". The answer then is simple; the mechanisms involved both in predicting where a ball would fall are similar to those for learning complex grammatical structures. This quickly faced the practical problems we have seen; the models either did not learn the 'right way' or they collapsed after a few hundred examples. Bayesian cognitive scientists have taken up this challenge turning it on its head. The learning mechanisms are now irrelevant and what matters is how the current input matches our prior expectations: "We do not need to know much if we understand where our data comes from". This faced a different sort of criticism. The mechanisms are now too detached, or the *prior* expectations are too artificial to match the experimental paradigm.

Perhaps a starting point to understand the above questions would be to take a step back and appreciate Hinton's quote at the start of this thesis. We do not need to know much coming to this world, not because we possess mechanisms that would quickly enable us to learn or because prior expectations guide our inferences, but because the inevitable contact with the world provides us with an input rich in statistics, regularities and distributional properties. While it might not contain *all* the information, it contains enough to get us started; to bootstrap our expectations and to learn *how to learn*; the rest is up to us.

Abbreviations

2AFC Two alternative forced choice task. 29, 129, 157, 171, 182

AGL Artificial Grammar Learning. 25

AIC Akaike Information Criterion. 74, 75, 191, 193

AN Association Norms. 20, 21, 72, 73, 108, 109

ANOVA Analysis of Variance. 86, 87, 257

ARC Average Radius of Co-Occurrence. 101

ARI Adjusted Rand Index. 162, 163

BEAGLE Bound Encoding of the Aggregate Language Environment. 46–48, 56, 60, 61, 72, 73, 75, 76, 78, 79, 83–86, 91, 107–110, 116, 268, 269

BLP British Lexicon Project. 101, 110, 111, 114

BNC British National Corpus. 55, 92, 102, 106, 110, 115, 166

BO Bayesian Optimization. 64, 65

BOW The Sinica Bilingual Ontological WordNet. 133

BPA Backward phrasal associate. 77, 79, 81

BTP Backward transitional probabilities. 189, 192

COALS Correlated Occurrence Analogue to Lexical Semantics. 46, 56, 72, 73, 75, 78, 80, 83, 86, 87, 107–110, 268, 269

COW Chinese Open WordNet. 133

CWN Taiwan University WordNet. 133

DSM Distributional Semantics Model. 55, 62, 64–66, 71–73, 85, 89, 90, 96, 103, 104, 108–110, 112, 116, 117, 132, 134, 135, 156, 204, 269

ERP Event-related potential. 21, 122

EUROPARL The Europarl parallel corpus. 187, 189, 190

FAS Forward Association Strength. 69

FN Feature Norms. 22

FPA Forward Phrasal Associate. 77, 79

FSG Finite-state grammar. 4, 7, 9, 23, 253, 254

FTP Forward Transitional Probabilities. 189, 192

HAL Hyperspace Analogue to Language. 44–46, 51, 56, 60–62, 72–76, 78–80, 83–87, 91–93, 98, 107–110, 211, 268, 269

IPA International Phonetic Alphabet. 179, 270

LDRT Lexical decision reaction time. 96–99, 101, 102, 106, 109, 113, 114, 260

LMM Linear mixed-effects regression model. 191, 193

LSA Latent Semantic Analysis. 43, 44, 46, 47, 55, 56, 58, 60–62, 69, 72, 73, 75, 76, 78–80, 83–87, 89, 91–93, 107–110, 211, 268

MDS Multidimensional Scaling. 130, 171

MWE Multiword expression. 54, 55

NCOUNT Number of words within a radius. 100, 101

NE Neural Embeddings. 79

NLP Natural Language Processing. 23, 88, 195, 210

NN Neural Network. 90

OLS Ordinary Least Squares. 191, 193

PCA Principal Components Analysis. 37, 130, 166, 167, 171, 183

PCFG Probabilistic Context Free Grammar. 35

PDP Parallel Distributed Processing. 12

POS Part of speech. 187

RI Random Indexing. 78

RT Reaction Time. 65–67, 70, 71, 98, 101

SEW Southeast University WordNet. 133

SIL Semantic Implicit Learning. 32, 33, 58, 111, 118, 119, 122, 150, 158, 171, 172, 186, 197, 198, 211, 265

SN Semantic Neighbourhood. 98, 101

SND Semantic Neighbourhood Density. 99, 107–111, 113

SOA Stimulus Onset Asynchrony. 60, 69, 70, 74, 82, 84, 89–91, 106, 116, 257

SPP Semantic Priming Project. 60, 65, 67–70, 72–74, 77, 78, 80, 81, 89, 90, 101, 102, 106, 108–111, 114, 257, 258

SVD Singular Value Decomposition. 42–44, 46, 49, 52, 69, 107, 131, 211, 258, 259, 268

TASA The Touchstone Applied Science Associates corpus. 55, 60, 92

TTR Type-Token Ratio. 104

VIF Variance Inflation Factor. 110

WALS World Atlas of Language Structures. 195

Notation

This dissertation has borrowed elements from different disciplines, namely statistics, machine learning, mathematics and psychology. The –quite hard to overcome– problem is that each of these disciplines follows its own conventions for the same things making it quite hard for us to be consistent with all of them. For example, cognitive psychologists use sum notation in order to find the net input to a particular unit in a neural network, whereas we use the dot product of the two matrices to obtain the vector of net inputs for the entire layer. We generally follow Murphy (2012) in our notation. Some of the symbols might not need any further explanation, but in order to be consistent we think it is better to cover all of the to avoid potential confusion. Note that it might be possible that we use the same symbol in different situations, in which case the meaning should be derived from context.

Symbol	Meaning
General math	
$\mathbf{x} \odot \mathbf{y}$	The Hadamard product (i.e., element-wise multiplication between $\mathbf{x}$ and $\mathbf{y}$)
$\alpha \wedge \beta$	logical AND
$\alpha \vee \alpha$	logical OR
$\neg\alpha$	logical NOT
∞	Infinity
$\rightarrow$	'Tends towards'
$\propto$	Proportional to
$\lvert x \rvert$	Absolute value
$\lvert \mathcal{S} \rvert$	Cardinality (i.e., size) of a set
∇	Jacobian vector (i.e., vector of first derivatives)
$f^{\star}$	The true underlying function
$\triangleq$	Defined as
$\Theta(\cdot)$	Grows as fast as
$O(\cdot)$	Grows no faster than

Continued on next page

Table 7.4 – continued from previous page

Symbol	Meaning
$\mathbb{R}$	Set of real numbers
$\approx$	Approximately equal to
$\arg\max_x f(x)$	the value x that maximises f
$\mathcal{X}$	A set (e.g., $\mathcal{V}$ is the corpus vocabulary)
Linear Algebra	
$\mathbf{M}^\top$	Transpose of a matrix
$\mathbf{v}^\top$	Transpose of a vector
$\mathbf{0}$, $\mathbf{1}$	Vector of zeros or ones
$\|x\|$	Euclidean norm
$\mathbf{M}_{*j}$	jth column of a matrix
$\mathbf{M}_{i*}$	ith row of a matrix
$\mathbf{M}_{ij}$	Element of matrix
$\mathbf{v}_i$	Element of vector
$\text{vec}(\mathbf{M})$	Vectorised form of the matrix
Probability	
$X \sim p$	X is sampled from p
$\mathbb{KL}(p\|q)$	KL divergence from distribution p to q
μ	mean of a vector
$P(x)$	probability of x
$P(x\|y)$	probability of x given y

Bibliography

Abrams, M., & Reber, A. S. (1988). Implicit learning: Robustness in the face of psychiatric disorders. *Journal of Psycholinguistic Research*, *17*, 425–439.

Adger, D., & Harbour, D. (2007). Syntax and syncretisms of the person case constraint. *Syntax*, *10*, 2–37.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Alikaniotis, D., & Williams, J. N. (2015). A distributional semantics approach to implicit language learning. In V. Pirrelli, C. Marzi, & M. Ferro (Eds.), *Word Structure and Word Usage. Proceedings of the NetWordS Final Conference* (pp. 81–84).

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 715–725). Association for Computational Linguistics.

Altmann, G. T. M., Dienes, Z., & Goode, A. (1995). Modality independence of implicitly learned grammatical knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 899–912.

Ambinder, M. S., Wang, R. F., Crowell, J. A., Francis, G. K., & Brinkmann, P. (2009). Human four-dimensional spatial intuition in virtual reality. *Psychonomic Bulletin & Review*, *16*, 818–823.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*, 261–295.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Andrews, M., Frank, S., & Vigliocco, G. (2014). Reconciling embodied and distributional accounts of meaning in language. *Topics in Cognitive Science*, *6*, 359–370.

Andrews, M., Vigliocco, G., & Vinson, D. P. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*, 463–498.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*, 234–254.

Arthur, D., & Vassilvitskii, S. (2006). How slow is the k-means method? In *Proceedings of the Twenty-second Annual Symposium on Computational Geometry* SCG '06 (pp. 144–153). New York, NY, USA: ACM.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Baayen, R. H. (2001). *Word frequency distributions*. Boston, MA: Kluwer.

Baayen, R. H. (2010). Demythologizing the word frequency effect. *The Mental Lexicon*, *5*, 436–461.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*, 12–28.

Baddeley, A. D. (2003). Working memory and language: an overview. *Journal of Communication Disorders*, *36*, 189–208.

Bai, X., Yan, G., Liversedge, S. P., Zang, C., & Rayner, K. (2008). Reading spaced and unspaced Chinese text: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 1277–1287.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283–316.

Balota, D. A., & Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *12*, 336–345.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459.

Barclay, J. R., Bransford, J. D., & Franks, J. J. (1974). Comprehension and semantic flexibility. *Journal of Verbal Learning and Verbal Behavior*, *13*, 471–481.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, *43*, 209–226.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 238–247). Association for Computational Linguistics.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*, 211–227.

Barsalou, L. W. (1987). Are there static category representations in long-term memory. *Behavioral and Brain Sciences*, *9*, 651–652.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

Bates, E. A., & Elman, J. L. (1992). *Connectionism and the Study of Change*. Technical Report 9202 University of California, San Diego San Diego, CA.

Batterink, L. J., Oudiette, D., Reber, P. J., & Paller, K. A. (2014). Sleep facilitates learning a new linguistic rule. *Neuropsychologia*, *65*, 169–179.

Becker, C. A. (1979). Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 252–259.

Becker, C. A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory & Cognition*, *8*, 493–512.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*, 521–547.

Beltagy, I., Chau, C., Boleda, G., Garrette, D., Erk, K., & Mooney, R. (2013). Montague meets markov: Deep semantics with probabilistic logical form. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity* (pp. 11–21). Association for Computational Linguistics.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017a). The entropy of words—learnability and expressivity across more than 1000 languages. *Entropy*, *19*, 275.

Bentz, C., Alikaniotis, D., Samardžić, T., & Buttery, P. (2017b). Variation in word frequency distributions: Definitions, measures and implications for a corpus-based language typology. *Journal of Quantitative Linguistics*, (pp. 1–35).

Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLOS ONE*, *10*, e0128254.

Berlin, B., & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution*. Berkeley, CA: University of California Press.

Berzak, Y., Barbu, A., Harari, D., Katz, B., & Ullman, S. (2015). Do you see what i mean? visual resolution of linguistic ambiguities. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1477–1487). Association for Computational Linguistics.

Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and Brain Sciences*, *7*, 173.

Bod, R., Hay, J. F., & Jannedy, S. (Eds.) (2001). *Probabilistic linguistics*. Cambridge, MA: The MIT Press.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*, 389–414.

Braine, M. D. S. (1963). On learning the grammatical order of words. *Psychological Review*, *70*, 323–348.

Braine, M. D. S. (1965). The insufficiency of a finite state model for verbal reconstructive memory. *Psychonomic Science*, *2*, 291–292.

Braine, M. D. S. (1966). Learning the positions of words relative to a marker element. *Journal of Experimental Psychology*, *72*, 532–540.

Braine, M. D. S. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Lawrence Erlbaum.

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. New York: Oxford University Press.

Brand, M. (2006). Fast low-rank modifications of the thin singular value decomposition. *Linear Algebra and its Applications*, *415*, 20–30.

Bresnan, J., & Hay, J. (2008). Gradient grammar: An effect of animacy on the syntax of give in New Zealand and American English. *Lingua*, *118*, 245–259.

Brody, J., Gumperz, J. J., & Levinson, S. C. (1998). Rethinking linguistic relativity. *Language*, *74*, 638.

Brooks, P. J., Braine, M. D. S., Catalano, L., & Brody, R. E. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language*, *32*, 76–95.

Brown, R. W. (1957). Linguistic determinism and the part of speech. *The Journal of Abnormal and Social Psychology*, *55*, 1–5.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911.

Buchanan, L., Burgess, C., & Lund, K. (1996). Overcrowding in semantic neighbourhoods: Modelling deep dyslexia. *Brain and Cognition*, *32*, 111–114.

Buchanan, L., Hildebrandt, N., & MacKinnon, G. E. (1994). Phonological processing of nonwords by a deep dyslexic patient: A rowse is implicitly a rose. *Journal of Neurolinguistics*, *8*, 163–181.

Buchanan, L., Hildebrandt, N., & MacKinnon, G. E. (1999). Phonological Processing Acquired Deep Dyslexia Reexamined. In R. M. Klein, & P. A. McMullen (Eds.), *Converging methods for understanding reading and dyslexia*. The MIT Press.

Buchanan, L., Kiss, I., & Burgess, C. (2000). Phonological and semantic information in word and nonword reading in a deep dyslexic patient. *Brain and Cognition*, *43*, 65–68.

Buchanan, L., Westbury, C. F., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, *8*, 531–544.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, *32*, 13–47.

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich, & A. B. Markman (Eds.), *Cognitive Dynamics: Conceptual Change in Humans and Machines* (pp. 117–156). Hillsdale, NJ: Lawrence Erlbaum.

Bybee, J. L., & Hopper, P. J. (Eds.) (2001). *Frequency and the Emergence of Linguistic Structure*. John Benjamins Publishing Company.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE*, *5*, e10729.

Capitani, E., Laiacona, M., Mahon, B., & Caramazza, A. (2003). What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. *Cognitive Neuropsychology*, *20*, 213–261.

Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience*, *10*, 1–34.

Chao, Y. R. (1968). *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, *119*, 417–430.

Chen, W., Guo, X., Tang, J., Zhu, L., Yang, Z., & Dienes, Z. (2011). Unconscious structural knowledge of form-meaning connections. *Consciousness and Cognition*, *20*, 1751–1760.

Chiarello, C., Burgess, C., Richards, L., & Pollock, A. (1990). Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't… sometimes, some places. *Brain and language*, *38*, 75–104.

Chomsky, N. A. (1957). *Syntactic Structures*. The Hague: Mouton and Co.

Chomsky, N. A. (1977). *Essays on Form and Interpretation*. New York, NY: North-Holland.

Chomsky, N. A. (1986). *Barriers*. Cambridge, MA: The MIT Press.

Chomsky, N. A. (1995). *The Minimalist Program*. Number 28 in Current studies in linguistics. Cambridge, MA: The MIT Press.

Christiansen, M. H., & Chater, N. (2016a). *Creating Language*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Chater, N. (2016b). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*.

Chubala, C. M., Johns, B. T., Jamieson, R. K., & Mewhort, D. J. K. (2016). Applying an exemplar model to an implicit rule-learning task: Implicit learning of semantic structure. *The Quarterly Journal of Experimental Psychology*, *69*, 1049–1055.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS workshop on Deep Learning and Representation Learning*. Montreal, Canada.

Chwilla, D. J., & Kolk, H. H. J. (2002). Three-step priming in lexical decision. *Memory & Cognition*, *30*, 217–225.

Cibelli, E., Xu, Y., Austerweil, J. L., Griffiths, T. L., & Regier, T. (2016). The Sapir-Whorf Hypothesis and Probabilistic Inference: Evidence from the Domain of Color. *PLOS ONE*, *11*, e0158725.

Clark, A., & Karmiloff-Smith, A. (1993). The Cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind & Language*, *8*, 487–519.

Clark, S. (2015). Vector space models of lexical meaning. In S. Lappin, & C. Fox (Eds.), *Handbook of Contemporary Semantics*. Wiley-Blackwell.

Cleeremans, A. (1997). Principles for implicit learning. In *How Implicit Is Implicit Learning?* (pp. 195–234). Oxford University Press.

Cleeremans, A. (2014). Connecting conscious and unconscious processing. *Cognitive Science*, *38*, 1286–1315.

Cleeremans, A., & Dienes, Z. (2008). Computational models of implicit learning. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 396–421). Cambridge, UK: Cambridge University Press.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240–247.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. *Proceedings of the Twenty-Fifth international conference on Machine Learning*, (pp. 160–167).

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.

Conklin, K., & Schmitt, N. (2007). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, *29*, 72–89.

Corbett, G. G. (1991). *Gender*. Cambridge Textbooks in Linguistics. Cambridge, UK: Cambridge University Press.

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the unexpected: Event-related brain response to morphosyntactic violations. *Language and Cognitive Processes*, *13*, 21–58.

Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science*, *23*, 371–414.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge, UK: Cambridge University Press.

Crutch, S. J., & Warrington, E. K. (2005). Abstract and concrete concepts have structurally different representational frameworks. *Brain*, *128*, 615–627.

Cucerzan, S., & Yarowsky, D. (2003). Minimally supervised induction of grammatical gender. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we able to run any psycholinguistic experiments at all in 1990? *Cognition*, *10*, 65–70.

Davidson, R., & MacKinnon, J. G. (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica*, *49*, 781.

Deacon, D., Dynowska, A., Ritter, W., & Grose-Fifer, J. (2004). Repetition and semantic priming of nonwords: Implications for theories of N400 and word recognition. *Psychophysiology*, *41*, 60–74.

Deerwester, S. C., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

DeKeyser, R. M. (1995). Learning second language grammar rules. *Studies in Second Language Acquisition*, *17*, 379.

Dell, G. S., Reed, K. D., Adams, D. R., & Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1355–1367.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*, 1117–1121.

Demeester, T., Rocktäschel, T., & Riedel, S. (2016). Lifted rule injection for relation embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1389–1399). Association for Computational Linguistics.

Denny, J. P., & Creider, C. A. (1976). The semantics of noun classes in Proto-Bantu. *Studies in African Linguistics*, *7*, 1–30.

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2013). The centre for speech, language and the brain (CSLB) concept property norms. *Behavior Research Methods*, *46*, 1119–1127.

Deyne, S. D., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General*, *145*, 1228–1254.

Deyne, S. D., Navarro, D. J., & Storms, G. (2012). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, *45*, 480–498.

Deyne, S. D., & Storms, G. (2008). Word associations: Norms for 1,424 dutch words in a continuous task. *Behavior Research Methods*, *40*, 198–205.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, *25*, 108–127.

Dikker, S., & Pylkkänen, L. (2013). Predicting language: MEG evidence for lexical preactivation. *Brain and Language*, *127*, 55–64.

Dixon, R. M. W. (1972). *The Dyirbal Language of North Queensland*. Cambridge, UK: Cambridge University Press.

Domangue, T. J., Mathews, R. C., Sun, R., Roussel, L. G., & Guidry, C. E. (2004). Effects of model-based and memory-based processing on speed and accuracy of grammar string generation. *Journal of Experimental Psychology*, *30*, 1002–1011.

Dryer, M. S., & Haspelmath, M. (Eds.) (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*, 2121–2159.

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, *6*, 126–155.

Ellis, N. C. (1994). Implicit and explicit processes in language acquisition. an introduction. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 1–32). San Diego, CA: Academic Press.

Ellis, N. C. (2002). Frequency Effects in Language Processing. *Studies in Second Language Acquisition*, *24*, 143–188.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit knowledge. *Studies in Second Language Acquisition*, *27*, 305–352.

Ellis, N. C. (2015). Implicit AND explicit language learning. In P. Rebuschat (Ed.), *Implicit and Explicit Learning of Languages* chapter 1. (pp. 3–24). Amsterdam: Johns Benjamins Publishing Company.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, *14*, 179–211.

Elman, J. L. (1991). Distributed Representations, Simple Recurrent Networks, And Grammatical Structure. *Machine Learning*, *7*, 195–225.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: The MIT Press.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*, 211–245.

Ettinger, A., & Linzen, T. (2016). Evaluating vector space models using human semantic priming results. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (pp. 72–77). Association for Computational Linguistics.

Evans, N. (1997). Role or cast? noun incorporation and complex predicates in Mayali. In J. B. A. Alsina, & P. Sells (Eds.), *Complex Predicates* (pp. 397–430). CSLI, Stanford.

Evans, N. (2003). *Bininj Gun-wok: a pan-dialectal grammar of Mayali, Kunwinjku and Kune*. Canberra: Pacific Linguistics.

Eysenck, M. (2008). *Fundamentals of Psychology*. Informa UK Limited.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*, 491–505.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

Ferrand, L., & New, B. (2003). Semantic and associative priming in the mental lexicon. In P. Bonin (Ed.), *Mental lexicon: Some words to talk about words* (pp. 25–43). Hauppage, NY: Nova Science Publishers.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The french lexicon project: Lexical decision data for 38,840 french words and 38,840 pseudowords. *Behavior Research Methods*, *42*, 488–496.

Fine, A. B., & Jaeger, T. F. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, *37*, 578–591.

Firth, J. R. (1935). The Technique of Semantics. *Transactions of the Philological Society*, *34*, 36–73.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in linguistic analysis* (pp. 1–32). Oxford, UK: Philological Society.

Flores d'Arcais, G. B., Schreuder, R., & Glazenborg, G. (1985). Semantic activation during recognition of referential words. *Psychological Research*, *47*, 39–49.

Fodor, J. A. (1975). *The language of thought*. New York, NY: Harper & Row.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: The MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.

Foucart, A., & Frenck-Mestre, C. (2010). Grammatical gender processing in L2: Electrophysiological evidence of the effect of L1–L2 syntactic similarity. *Bilingualism: Language and Cognition*, *14*, 379–399.

Foucart, A., & Frenck-Mestre, C. (2012). Can late L2 learners acquire new grammatical features? evidence from ERPs and eye-tracking. *Journal of Memory and Language*, *66*, 226–248.

Frank, L. S., Otten, J. L., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts n400 amplitude during reading. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 878–883). Association for Computational Linguistics.

Frank, M. C., Goldwater, S., Mansinghka, V., Griffiths, T. L., & Tenenbaum, J. B. (2007). Modeling Human Performance in Statistical Word Segmentation. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, *1*.

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*, 360–371.

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language*, *39*, 218–245.

Ganesan, K., Zhai, C., & Viegas, E. (2012). Micropinion generation: An unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st International Conference on World Wide Web 2012 (WWW '12)*.

Gao, M. Y., & Malt, B. C. (2009). Mental representation and cognitive consequences of Chinese individual classifiers. *Language and Cognitive Processes*, *24*, 1124–1179.

Gao, Y., Kontoyiannis, I., & Bienenstock, E. (2008). Estimating the entropy of binary time series: Methodology, some theory and a simulation study. *Entropy*, *10*, 71–99.

Garrette, D., Erk, K., & Mooney, R. (2014). A formal approach to linking logical form and vector-space lexical semantics. In *Text, Speech and Language Technology* (pp. 27–48). Springer Science and Business Media.

Gazdar, G., Klein, E., Pullum, G., & Sag, I. (1985). *Generalized Phrase Structure Grammar*. Cambridge, MA: Harvard University Press.

Gelman, R. (1990). First principles organize attention to and learning about relevant data: Number and the animate-inanimate distinction as examples. *Cognitive Science*, *14*, 79–106.

Gillon Dowens, M., Guo, T., Guo, J., Barber, H. A., & Carreiras, M. (2011). Gender and number processing in Chinese learners of Spanish – evidence from event related potentials. *Neuropsychologia*, *49*, 1651–1659.

Gillon Dowens, M., Vergara, M., Barber, H. A., & Carreiras, M. (2010). Morphosyntactic processing in late second-language learners. *Journal of Cognitive Neuroscience*, *22*, 1870–1887.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, *9*, 249–256.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*, (pp. 673–680).

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21–54.

Goldwater, S., & Johnson, M. (2005). Representational bias in unsupervised learning of syllable structure. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)* (pp. 112–119). Association for Computational Linguistics.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, MA: The MIT Press. `http://www.deeplearningbook.org`.

Gorder, P. F. (2006). Neural Networks Show New Promise for Machine Vision. *Computing in Science and Engineering*, (pp. 4–8).

Goswami, U. (2008). *Cognitive Development: The Learning Brain*. Hove and New York: Psychology Press.

Goujon, A. (2011). Categorical implicit learning in real-world scenes: Evidence from contextual cueing. *The Quarterly Journal of Experimental Psychology*, *64*, 920–941.

Goujon, A., Didierjean, A., & Marmèche, E. (2009). Semantic contextual cuing and visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 50–71.

Goujon, A., Didierjean, A., & Thorpe, S. (2015). Investigating implicit statistical learning mechanisms through contextual cueing. *Trends in Cognitive Sciences*, *19*, 524–533.

Grainger, J., O'Regan, J. K., Jacobs, A. M., & Segui, J. (1989). On the role of competing word units in visual word recognition: The neighborhood frequency effect. *Perception & Psychophysics*, *45*, 189–195.

Graves, A., Mohamed, A., & Hinton, G. E. (2013). Speech Recognition with Deep Recurrent Neural Networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 6645–6649).

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural turing machines. *CoRR*, *abs/1410.5401v2*.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin*, *138*, 415–422.

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

Griffiths, T. L., & Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition*, *103*, 180–226.

de Groot, A. M. B. (1983). The range of automatic spreading activation in word priming. *Journal of Verbal Learning and Verbal Behavior*, *22*, 417–436.

Gutmann, M. U., & Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, *13*, 307–361.

H., B. R., Piepenbrock, R., & Guilikers, L. (1995). *The CELEX Lexical Database (CD-ROM) (Version 2.5)*. Philadelphia, PA: Linguistic Data Consortium.

Hagoort, P., Brown, C. M., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, *8*, 439–483.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Harabagiu, S. (Ed.) (1998). *Proceedings of the Wordshop on the Usage of WordNet in Natural Language Processing Systems*. Montréal, Canada: Association for Computational Linguistics.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*, 491–528.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720.

Harris, Z. (1954). Distributional structure. *Word*, *10*, 146–162.

Hashimoto, T., Alvarez-Melis, D., & Jaakkola, T. (2016). Word embeddings as metric recovery in semantic spaces. *Transactions of the Association of Computational Linguistics – Volume 4, Issue 1*, (pp. 273–286).

Hauk, O., Davis, M., Ford, M., Pulvermüller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage*, *30*, 1383–1400.

Hawkinson, A., & Hyman, L. (1974). Hierarchies of natural topic in Shona. *Studies in African Linguistics*, *5*, 147–170.

Hecht-Nielsen, R. (1994). Context vectors; general purpose approximate meaning representations self-organized from raw data. In J. M. Zurada, R. J. Marks, & C. J. Robinson (Eds.), *Computational Intelligence: imitating life*. IEEE Press.

Heider, E. R. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, *93*, 10–20.

Heider, E. R., & Olivier, D. C. (1972). The structure of the color space in naming and memory for two languages. *Cognitive Psychology*, *3*, 337–354.

van Hell, J. G., & Tokowicz, N. (2010). Event-related brain potentials and second language learning: syntactic processing in late L2 learners at different L2 proficiency levels. *Second Language Research*, *26*, 43–74.

Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada* (pp. 1693–1701).

Heyman, T., Bruninx, A., Hutchison, K. A., & Storms, G. (2018). The (un)reliability of item-level semantic priming effects. *Behavior Research Methods*, (pp. 1–11).

Hill, F., & Korhonen, A. (2014). Learning abstract concept embeddings from multi-modal data: Since you probably can't see what i mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 255–265). Association for Computational Linguistics.

Hill, F., Korhonen, A., & Bentz, C. (2014). A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive Science*, *38*, 162–177.

Hills, T. (2012). The company that words keep: comparing the statistical structure of child-versus adult-directed language. *Journal of Child Language*, *40*, 586–604.

Hinrichs, L., & Szmrecsányi, B. (2007). Recent changes in the function and frequency of standard english genitive constructions: a multivariate analysis of tagged corpora. *English Language and Linguistics*, *11*.

Hinton, G. E. (2014). Where do features come from? *Cognitive Science*, *38*, 1078–1101.

Hinton, G. E., Dayan, P., Frey, B., & Neal, R. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science*, *268*, 1158–1161.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*, 504–507.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, *abs/1207.0580v1*.

Hinton, L., Nichols, J., & Ohala, J. (Eds.) (1994). *Sound Symbolism*. Cambridge, UK: Cambridge University Press.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*, 1735–1780.

Hohwy, J. (2013). *The Predictive Mind*. Oxford, UK: Oxford University Press.

Hopkins, J. D., & Schwanenflugel, P. J. (1993). The psychology of word meanings. *Language*, *69*, 222.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*, 359–366.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417–441.

Howard, M. W., & Kahana, M. J. (2002). When Does Semantic Similarity Help Episodic Retrieval? *Journal of Memory and Language*, *46*, 85–98.

Huang, C.-R., Chang, R.-Y., & Lee, H.-P. (2004). Sinica bow (bilingual ontological wordnet): Integration of bilingual wordnet and sumo. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA).

Huang, C.-R., Hsieh, S.-K., Hong, J.-F., Chen, Y.-Z., Su, I.-L., Chen, Y.-X., & Huang, S.-W. (2010). Chinese WordNet: Design and implementation of a cross-lingual knowledge processing infrastructure. *Journal of Chinese Information Processing*, *24*, 14–23.

Huang, C.-R., Tseng, E. I. J., Tsai, D. B. S., & Murphy, B. (2003). Cross-lingual portability of semantic relations: Bootstrapping Chinese WordNet with English WordNet relations. *Language and Linguistics*, *4*, 509–532.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220–264.

Hutchison, K. A. (2003a). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review*, *10*, 785–813.

Hutchison, K. a. (2003b). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic bulletin & review*, *10*, 785–813.

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology*, *61*, 1036–1066.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., Yap, M. J., Bengson, J. J., Niemeyer, D., & Buchanan, E. (2013). The Semantic Priming Project. *Behavior Research Methods*, *45*, 1099–1114.

Hutchison, K. A., Neely, J. H., & Johnson, J. D. (2001). With great expectations, can two 'wrongs' prime a 'right'? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*, 1451–1463.

Iacobacci, I., Pilehvar, T. M., & Navigli, R. (2016). Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 897–907). Association for Computational Linguistics.

Imai, M., Schalk, L., Saalbach, H., & Okada, H. (2014). All Giraffes Have Female-Specific Properties: Influence of Grammatical Gender on Deductive Reasoning About Sex-Specific Properties in German Speakers. *Cognitive Science*, *38*, 514–536.

Jackendoff, R. (2002). *Foundations of Language*. Oxford: Oxford University Press.

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2016). An R^2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, *44*, 1086–1105.

Jiang, J. J., & Conrath, W. D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th Research on Computational Linguistics International Conference* (pp. 19–33). The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, *91*, 433–445.

Jiang, S. (2017). *The semantics of Chinese classifiers and linguistic relativity*. Routledge studies in Chinese linguistics. London: Routledge.

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences*, *96*, 7592–7597.

Jones, J., & Pashler, H. (2007). Is the mind inherently forward looking? comparing prediction and retrodiction. *Psychonomic Bulletin & Review*, *14*, 295–300.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*, 534–552.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1–37.

Jordan, M. I. (1986). An introduction to linear algebra in parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* chapter 9. (pp. 365–422). Cambridge, MA: MIT Press volume 1: Foundations.

Kanerva, P. (2009). Hyperdimensional computing: An introduction to computing in distributed representation with high dimensional random vectors. *Cognitive Computation*, *1*, 139–159.

Kanerva, P., Kristoferson, J., & Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (p. 1036). Mahwah, NJ volume Erlbaum.

Kapatsinski, V., & Radicke, J. (2009). Frequency and the emergence of prefabs: Evidence from monitoring. In R. Corrigan, E. A. Moravcsik, H. Ouali, & K. Wheatley (Eds.), *Formulaic Language: Volume 2. Acquisition, loss, psychological reality, and functional explanations* (pp. 499–520). John Benjamins Publishing Company volume 83 of *Typological Studies in Language*.

Kaplan, R. M. (2003). Syntax. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics* chapter 4. (pp. 70–90). Oxford, UK: Oxford University Press.

Katz, J. J. (1980). Chomsky on meaning. *Language*, *56*, 1.

Keil, F. C. (1979). *Semantic and Conceptual Development*. Cambridge, MA: Harvard University Press.

Kelly, M. H. (1992). Using sound to solve syntactic problems: the role of phonology in grammatical category assignments. *Psychological Review*, *99*, 349–364.

Kerstens, J. (1993). *The Syntax of Number, Person and Gender*. Walter de Gruyter GmbH.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology*, *1*, 1–15.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior research methods*, *44*, 287–304.

Kiela, D., Bulat, L., & Clark, S. (2015a). Grounding semantics in olfactory perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 231–236). Association for Computational Linguistics.

Kiela, D., Bulat, L., Verő, A. L., & Clark, S. (2016). Virtual embodiment: A scalable long-term strategy for artificial intelligence research. *CoRR*, *abs/1610.07432v1*.

Kiela, D., & Clark, S. (2015). Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2461–2470). Association for Computational Linguistics.

Kiela, D., Hill, F., & Clark, S. (2015b). Specializing word embeddings for similarity or relatedness. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2044–2048). Association for Computational Linguistics.

Kintsch, E., Caccamise, D., Franzke, M., Johnson, M., & Dooley, S. (2007). Summary street: Computer-guided summary writing. In T. K. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 263–278). Mahwah, NJ: Erlbaum.

Kintsch, W., & Mangalath, P. (2011). The construction of meaning. *Topics in Cognitive Science*, *3*, 346–370.

Knowlton, B. J., Ramus, S. J., & Squire, L. R. (1992). Intact artificial grammar learning in amnesia: Dissociation of classification learning and explicit memory for specific instances. *Psychological Science*, *3*, 172–179.

Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 79–91.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit* (pp. 79–86). volume 5.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59–69.

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*, 211–240.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. *Assessment in Education*, *10*, 295–308.

Lany, J., & Saffran, J. R. (2010). From statistics to meaning: Infants' acquisition of lexical categories. *Psychological Science*, *21*, 284–291.

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language*, *61*, 326–338.

Lau, E. F., Weber, K., Gramfort, A., Hämäläinen, M. S., & Kuperberg, G. R. (2014). Spatiotemporal signatures of lexical–semantic prediction. *Cerebral Cortex*, *26*, 1377–1387.

Lazaridou, A., Marelli, M., & Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, *41*, 677–705.

Leacock, C., & Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database* (pp. 265–283). Cambridge, MA: The MIT Press.

LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, *521*, 436–444.

Lemaire, B., & Denhière, G. (2006). Effects of high-order co-occurrences on word semantic similarity. *Current Psychology Letters: Behaviour, Brain & Cognition*, *18*.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica*, *20*, 1–31.

Leung, J. H. C., & Williams, J. N. (2012). Constraints on Implicit Learning of Grammatical Form-Meaning Connections. *Language Learning*, *62*, 634–662.

Leung, J. H. C., & Williams, J. N. (2013). Prior Linguistic Knowledge Influences Implicit Language Learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 2866–2871). Austin, TX: Cognitive Science Society.

Leung, J. H. C., & Williams, J. N. (2014). Crosslinguistic Differences in Implicit Language Learning. *Studies in Second Language Acquisition*, *36*, 733–755.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 1055–1065). Association for Computational Linguistics.

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native spanish speakers. *Journal of Memory and Language*, *63*, 447–464.

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: a Functional Reference Grammar*. Berkeley, CA: University of California Press.

Liang, F. M. (1983). *Word Hy-phen-a-tion by Com-pu-ter*. Ph.D. thesis Stanford University Department of Computer Science.

Lieberman, M. D. (2000). Intuition: A social cognitive neuroscience approach. *Psychological Bulletin*, *126*, 109–137.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*.

Lison, P., & Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016)*. European Languages Resources Association (ELRA).

Livesay, K., & Burgess, C. (1998). Mediated priming in high-dimensional semantic space: No effect of direct semantic relationships or co-occurrence. *Brain & Cognition*, *37*, 102–105.

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review*, *7*, 618–630.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*, *19*, 1–36.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*, 203–208.

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In J. D. Moore, & J. F. Lehman (Eds.), *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 660–665). Mahwah, NJ: Lawrence Erlbaum Associates.

Lund, K., Burgess, C., & Audet, C. (1996). Dissociating semantic and associative word relationships using high-dimensional space. In G. W. Cottrell (Ed.), *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (pp. 603–608). Mahwah, NJ: Lawrence Erlbaum Associates.

Lyons, J. (1977). *Semantics*. Cambridge, UK: Cambridge University Press.

van der Maaten, L., & Hinton, G. E. (2011). Visualizing non-metric similarities in multiple maps. *Machine Learning*, *87*, 33–55.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Manning, C., & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, *11*, F9–F16.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt and Co., Inc.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*, 71–102.

Mathews, R. C., Buss, R. R., Chinn, R., & Stanley, W. B. (1988). The role of explicit and implicit learning processes in concept discovery. *The Quarterly Journal of Experimental Psychology Section A*, *40*, 135–165.

Mathews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J. R., & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1083–1100.

McAndrews, M. P., & Moscovitch, M. (1985). Rule-based and exemplar-based classification in artificial grammar learning. *Memory & Cognition*, *13*, 469–475.

McCauley, S. M., & Christiansen, M. H. (2011). Learning Simple Statistics for Language Comprehension and Production : The CAPPUCCINO Model Simulation 1 : Modeling Comprehension and. *Production*, (pp. 1619–1624).

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, *102*, 419–457.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological Review*, *88*, 375–407.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*, 295–322.

McKoon, G., & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*, 1155–1172.

McNamara, T. P. (1992). Theories of priming: I. Associative dstance and lag. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*, 1173–1190.

McNamara, T. P., & Altarriba, J. (1988). Depth of spreading activation revisited: Semantic mediated priming occurs in lexical decisions. *Journal of Memory and Language*, *27*, 545–559.

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*, 547–559.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*, 99–130.

Mehler, J., Bertoncini, J., Dupoux, E., & Pallier, C. (1996). The role of suprasegmentals in speech perception and acquisition. In T. Otake, & A. Cutler (Eds.), *Phonological Structure and Language Processing* Speech Research 12 chapter 6. (pp. 145–170). The Hague: De Gruyter Mouton.

Meulemans, T., & der Linden, M. V. (1997). Associative chunk strength in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1007–1028.

Meulman, N., Wieling, M., Sprenger, S. A., Stowe, L. A., & Schmid, M. S. (2015). Age effects in L2 grammar processing as revealed by ERPs and how (not) to study them. *PLOS ONE*, *10*, e0143328.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology: General*, *90*, 227–234.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013a). Distributed Representations of Words and Phrases and their Compositionality. *Nips*, (pp. 1–9).

Mikolov, T., Corrado, G. S., Chen, K., & Dean, J. (2013b). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, (pp. 1–12).

Mikolov, T., Sutskever, I., Chen, K., & Corrado, G. S. (2013c). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* NIPS'13 (pp. 3111–3119).

Mikolov, T., Yih, W.-t., & Zweig, G. (2013d). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751). Association for Computational Linguistics.

Miller, G. A. (1958). Free recall of redundant strings of letters. *Journal of Experimental Psychology*, *56*, 485–491.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, *6*, 1–28.

Miller, G. A., & Fellbaum, C. (1992). WordNet and the organization of lexical memory. In S. M.L., & Y. M. (Eds.), *Intelligent Tutoring Systems for Foreign Language Learning* (pp. 89–102). Berlin: Springer Nature volume 80 of *NATO ASI Series (Series F: Computer and Systems Sciences)*.

Miozzo, M., Pulvermüller, F., & Hauk, O. (2015). Early parallel activation of semantics and phonology in picture naming: Evidence from a multiple linear regression MEG study. *Cerebral Cortex*, *25*, 3343–3355.

Mirman, D., & Magnuson, J. S. (2008). Attractor dynamics and semantic neighborhood density: Processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *34*, 65–79.

Mitchell, J., & Lapata, M. (2008). Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT* (pp. 236–244). Association for Computational Linguistics.

Mitchell, J., & Lapata, M. (2009). Language models based on semantic composition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 430–439). Association for Computational Linguistics.

Mitchell, J., & Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, *34*, 1388–1429.

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*, 1191–1195.

Mithun, M. (1986). The convergence of noun classification systems. In C. Craig (Ed.), *Noun classes and categorization* (pp. 379–397). Philadelphia, PA: John Benjamins.

Monaghan, P., Chang, Y.-N., Welbourne, S., & Brysbaert, M. (2017). Exploring the relations between word frequency, language exposure, and bilingualism in a computational model of reading. *Journal of Memory and Language*, *93*, 1–21.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*, 143–182.

Monaghan, P., Christiansen, M. H., Farmer, T. A., & Fitneva, S. A. (2012). Measures of phonological typicality. In *Benjamins Current Topics* (pp. 13–31). John Benjamins Publishing Company.

Monaghan, P., Lupyan, G., & Christiansen, M. H. (2014a). The Systematicity of the Sign: Modeling Activation of Semantic Attributes from Nonwords. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014b). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*, 1–12.

Montague, R. (1970). English as a formal language. In B. Visentini (Ed.), *Linguaggi nella societa e nella tecnica* (pp. 188–221). Edizioni di Communita.

Morolong, M., & Hyman, L. (1977). Animacy, objects and clitics in Sesotho. *Studies in African Linguistics*, *8*, 199–217.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*, 165–178.

Moscoso del Prado Martín, F., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: an information theoretical perspective on morphological processing. *Cognition*, *94*, 1–18.

Moss, H. E., McCormick, S. F., & Tyler, L. K. (1997). The Time Course of Activation of Semantic Information during Spoken Word Recognition. *Language and Cognitive Processes*, *12*, 695–732.

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 863–883.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*, 609–626.

Murdock, B. B. (1992). Item and associative information in a distributed memory model. *Journal of Mathematical Psychology*, *36*, 68–98.

Murdock, B. B. (1993). Derivations for the chunking model. *Journal of Mathematical Psychology*, *37*, 421–445.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive communication and machine learning series. Cambridge, MA: The MIT Press.

Nastase, V., & Popescu, M. (2009). What's in a name? in some languages, grammatical gender. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 1368–1377). Association for Computational Linguistics.

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*, 226–254.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner, & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264–336). Hillsdale, NJ: Erlbaum.

Neil, G. J., & Higham, P. A. (2012). Implicit learning of conjunctive rule sets: An alternative to artificial grammars. *Consciousness and Cognition*, *21*, 1393–1400.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*, 402–407.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Nieuwenhuis, I. L. C., Folia, V., Forkstam, C., Jensen, O., & Petersson, K. M. (2013). Sleep Promotes the Extraction of Grammatical Rules. *PLoS ONE*, *8*.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nuckolls, J. B. (2010). The sound-symbolic expression of animacy in amazonian ecuador. *Diversity*, *2*, 353–369.

Ó Séaghdha, D. (2007). Annotating and learning compound noun semantics. In *Proceedings of the ACL 2007 Student Research Workshop* (pp. 73–78). Association for Computational Linguistics.

de Oliveira, M., & Levkowitz, H. (2003). From visual data exploration to visual data mining: A survey. *IEEE Trans. Visual. Comput. Graphics*, *9*, 378–394.

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, *126*, 268–284.

Opitz, B., & Hofmann, J. (2015). Concurrence of rule- and similarity-based mechanisms in artificial grammar learning. *Cognitive Psychology*, *77*, 77–99.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: The MIT Press.

Ostrin, R. K., & Tyler, L. K. (1993). Automatic access to lexical semantics in aphasia: Evidence from semantic and associative priming. *Brain and language*, *45*, 147–159.

Ouyang, L., Boroditsky, L., & Frank, M. C. (2016). Semantic coherence facilitates distributional learning. *Cognitive Science*, *41*, 855–884.

Paciorek, A. (2013). *Implicit learning of semantic preferences*. Ph.D. thesis University of Cambridge Cambridge, UK.

Paciorek, A., & Williams, J. N. (2015). Semantic generalization in implicit language learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 989–1002.

Pacton, S., Perruchet, P., Fayol, M., & Cleeremans, A. (2001). Implicit learning out of the lab: The case of orthographic regularities. *Journal of Experimental Psychology: General*, *130*, 401–426.

Pacton, S., Sobaco, A., & Perruchet, P. (2015). Is an attention-based associative account of adjacent and nonadjacent dependency learning valid? *Acta Psychologica*, *157*, 195–199.

Paivio, A. (1971). *Imagery and verbal processes*. New York, NY: Holt, Rinehart & Winston.

Partee, B. H. (2014). A brief history of the syntax-semantics interface in western formal linguistics. *Sematnic-Syntax Interface*, *1*, 1–21.

Pastorino Campos, C. A. (2017). Exploring statistical learning of meaning-based regularities in children. Unpublished Manuscript.

Pecher, D., Zeelenberg, R., & Raaijmakers, J. (1998). Does pizza prime coin? Perceptual priming in lexical decision and pronunciation. *Journal of Memory and Language*, *38*, 401–418.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*, 244–247.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, *36*, 1299–1305.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General*, *119*, 264–275.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*, 233–238.

Perruchet, P., & Rey, A. (2005). Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? *Psychonomic bulletin & review*, *12*, 307–313.

Perruchet, P., & Vinter, A. (1998). Feature creation as a byproduct of attentional processing. *Behavioral and Brain Sciences*, *21*, 33–34.

Perruchet, P., & Vinter, A. (2002). The self-organizing consciousness. *Behavioral and Brain Sciences*, *25*.

Perruchet, P., Vinter, A., Pacteau, C., & Gallego, J. (2002). The formation of structurally relevant units in artificial grammar learning. *The Quarterly Journal of Experimental Psychology Section A*, *55*, 485–503.

Petig, W. E., Hammer, A. E., & Durrell, M. (1993). Hammer's German grammar and usage. *Die Unterrichtspraxis / Teaching German*, *26*, 115.

Pilehvar, T. M., Jurgens, D., & Navigli, R. (2013). Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1341–1351). Association for Computational Linguistics.

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*.

Plate, T. A. (2003). Holographic reduced representations. In *CSLI Lecture Notes* 150. Stanford, CA: CSLI Publications.

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. In J. D. Moore, & J. F. Lehman (Eds.), *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 37–42). Mahwah, NJ: Lawrence Erlbaum Associates.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes*, *12*, 765–806.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*, 377–500.

Polajnar, T., Rimell, L., & Clark, S. (2014). Evaluation of simple distributional compositional operations on longer texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Polajnar, T., Rimell, L., & Clark, S. (2015). An exploration of discourse-based sentence spaces for compositional distributional semantics. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics* (pp. 1–11). Association for Computational Linguistics.

Polinsky, M. (1996). The double object construction in spoken Eastern Armenian. *NSL. Linguistic Studies in the Non-Slavic Languages of the commonwealth of Independent States and the Baltic Republics*, *8*, 307–335.

Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 55–85). Hillsdale, NJ: Erlbaum.

Pulman, S. G. (1978). Shape classifiers and natural categories. *University of Essex Language Center Occasional Papers*, *20*, 35–57.

Qian, P., Qiu, X., & Huang, X. (2016). Investigating language universal and specific properties in word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1478–1488). Association for Computational Linguistics.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, *12*, 410–430.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing* (pp. 227–270). Cambridge, MA: MIT Press.

Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition*, *132*, 68–89.

Radovanović, M., Nanopoulos, A., & Ivanović, M. (2010). On the existence of obstinate results in vector space models. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10*. ACM Press.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*, 846–850.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, *97*, 285–308.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, *95*, 385–408.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*, 855–863.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of experimental psychology: General*, *118*, 219–235.

Reed, N., McLeod, P., & Dienes, Z. (2010). Implicit knowledge and motor skill: What people who know how to catch don't know. *Consciousness and Cognition*, *19*, 63–76.

Regier, T. (1997). The human semantic potential: Spatial language and constrained connectionism. *Computers & Mathematics with Applications*, *33*, 134.

Regier, T. (2003). Emergent constraints on word-learning: a computational perspective. *Trends in Cognitive Sciences*, *7*, 263–268.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *roceedings of the 14th international joint conference on Artificial intelligence - Volume 1 - IJCAI'95*, *1*, 6.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA: MIT Press.

Rohde, D., Gonnerman, L. M., & Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, *8*, 627–660.

Rosa, P. A. D., Catricalà, E., Vigliocco, G., & Cappa, S. F. (2010). Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 italian words. *Behavior Research Methods*, *42*, 1042–1048.

Rosch, E. (1978). Principles of categorization. In E. Rosch, & B. B. Lloyd (Eds.), *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Rosenblatt, M. (1958). A multi-dimensional prediction problem. *Arkiv för matematik*, *3*, 407–424.

Rothe, S., & Schütze, H. (2015). Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1793–1803). Association for Computational Linguistics.

Rouder, J. N., & Morey, R. D. (2012). Default bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877–903.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*, 627–633.

Ruge, G. (1992). Experiments on linguistically-based term associations. *Information Processing & Management*, *28*, 317–332.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986a). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: II. the contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*, 60–94.

Rumelhart, D. E., & McClelland, J. L. (1985). Levels indeed! A response to Broadbent. *Journal of Experimental Psychology: General*, *114*, 193–197.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* chapter 18. (pp. 216–271). Cambridge, MA: MIT Press volume 2: Psychological and Biological Models.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group (Eds.) (1986b). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. Cambridge, MA, USA: MIT Press.

Rumelhart, D. E., & Todd, P. (1993). Learning and connectionist representations. In D. E. Meyer, & S. Kornblum (Eds.), *Attention and Performance XIV: Synergies in experimental psychology, artificial intelligence and cognitive neuroscience* (pp. 3–30). Cambridge, MA: MIT Press.

Saalbach, H., & Imai, M. (2007). Scope of linguistic influence: Does a classifier system alter object concepts? *Journal of Experimental Psychology: General*, *136*, 485–501.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*, 1926–1928.

Sahlgren, M. (2006). *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis Department of Linguistics, Stockholm University.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, *20*.

Sahlgren, M., Holst, A., & Kanerva, P. (2006). Permutations as a means to encode order in word space. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1300–1305). Washington, DC.

de Saussure, F. (1916). *Course in General Linguistics*. New York, NY: McGraw-Hill.

Scellier, B., & Bengio, Y. (2017). Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, *11*, 24.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*, 486–492.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2016). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *372*.

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2015). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, *26*, 3–8.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Schmitt, B. H., & Zhang, S. (1998). Language structure and categorization: A study of classifiers in consumer cognition, judgment, and choice. *Journal of Consumer Research*, *25*, 108–122.

Schreuder, R., & Flores D' Arcais, G. B. (1992). Psycholinguistic issues in the lexical representation of meaning. In W. Marslen-Wilson (Ed.), *Lexical Representation and Process* chapter 14. (pp. 409–436). Cambridge, MA: MIT Press.

Schreuder, R., Flores d'Arcais, G. B., & Glazenborg, G. (1984). Effects of perceptual and conceptual similarity in semantic priming. *Psychological Research*, *45*, 339–354.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*, 2673–2681.

Sears, C. R., Hino, Y., & Lupker, S. J. (1995). Neighborhood frequency and neighborhood size effects in visual word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, *21*, 876–900.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–68.

Seidenberg, M. S., Tanenhaus, M. K., Leiman, J. M., & Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology*, *14*, 489–537.

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 592–608.

Shallice, T. (1988). Specialisation within the semantic system. *Cognitive Neuropsychology*, *5*, 133–142.

Shanks, D. R. (2016). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin & Review*, *early access*, 1–24.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 623–656.

Shaoul, C., Baayen, R. H., & Westbury, C. F. (2014). N-gram probability effects in a cloze task. *The Mental Lexicon*, *9*, 437–472.

Shaoul, C., & Westbury, C. F. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods*, *38*, 190–195.

Shaoul, C., & Westbury, C. F. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*, 393–413.

Shelton, J. R., & Martin, R. C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*, 1191–1210.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*, 390–398.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Siakaluk, P. D., Buchanan, L., & Westbury, C. F. (2003). The effect of semantic distance in yes/no and go/no-go semantic categorization tasks. *Memory & Cognition*, *31*, 100–113.

Smith, K. H. (1966). Grammatical intrusions in the recall of structured letter pairs: Mediated transfer or position learning? *Journal of Experimental Psychology*, *72*, 580–588.

Smith, K. H. (1969). Learning co-occurrence restrictions: Rule induction or rote learning? *Journal of Verbal Learning and Verbal Behavior*, *8*, 319–321.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 2951–2959). Curran Associates, Inc.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its applications in retrieval. *Journal of Documentation*, *28*, 11–21.

Srinivasan, M. (2010). Do classifiers predict differences in cognitive processing ? A study of nominal classification in Mandarin Chinese. *Language and Cognition*, *2*, 177–190.

Stanley, W. B., Mathews, R. C., Buss, R. R., & Kotler-Cope, S. (1989). Insight without awareness: On the interaction of verbalization, instruction and practice in a simulated process control task. *The Quarterly Journal of Experimental Psychology Section A*, *41*, 553–577.

Sun, R. (1997). Learning, action and consciousness: a hybrid approach toward modelling consciousness. *Neural Networks*, *10*, 1317–1331.

Sun, R. (2008). Introduction to computational cognitive modeling. In R. Sun (Ed.), *Cambridge Handbook of Computational Psychology* chapter 1. (pp. 3–19). New York: Cambridge University Press.

Sussna, M. J. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management (CIKM-93)* (pp. 67–74). Arlington, VA.

Swingley, D. (1999). Conditional probability and word discovery: A corpus analysis of speech to infant. In M. Hahn, & S. C. Stoness (Eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science society* (pp. 724–729). Mahwah, NJ: Erlbaum.

Sze, W. P., Liow, S. J. R., & Yap, M. J. (2013). The Chinese Lexicon Project: A repository of lexical decision behavioral responses for 2,500 chinese characters. *Behavior Research Methods*, *46*, 263–273.

Tang, D., Qin, B., Feng, X., & Liu, T. (2016). Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Association for Computational Linguistics.

Tenenbaum, J. B. (1999). *A Bayesian framework for concept learning*. Ph.D. thesis Massachusetts Institute of Technology Cambridge, MA.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

Tenenbaum, J. B., Perfors, A., & Regier, T. (2011a). The learnability of abstract syntactic principles. *Cognition*, *118*, 306–338.

Tenenbaum, J. B., Xu, F., Perfors, A., & Griffiths, T. L. (2011b). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*, 302–321.

Thiessen, E. D., & Pavlik, P. I. (2012). iMinerva: A mathematical model of distributional statistical learning. *Cognitive Science*, *37*, 310–343.

Thiessen, E. D., & Saffran, J. R. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development*, *3*, 73–100.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*, 273–282.

Till, R. E., Mross, E. F., & Kintsch, W. (1988). Time course of priming for associate and inference words in a discourse context. *Memory & Cognition*, *16*, 283–298.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika*, *17*, 401–419.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D., & Manning, C. (2005). A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Tsvetkov, Y., Faruqui, M., Ling, W., MacWhinney, B., & Dyer, C. (2016). Learning the curriculum with bayesian optimization for task-specific word representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 130–139). Association for Computational Linguistics.

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of Association for Computational Linguistics* (pp. 384–394). Association for Computational Linguistics.

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Tyler, L. K., Stamatakis, E. A., Jones, R. W., Bright, P., Acres, K., & Marslen-Wilson, W. D. (2004). Deficits for semantics and the irregular past tense: A causal relationship? *Journal of Cognitive Neuroscience*, *16*, 1159–1172.

Upadhyay, S., Faruqui, M., Dyer, C., & Roth, D. (2016). Cross-lingual models of word embeddings: An empirical comparison. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1661–1670). Association for Computational Linguistics.

van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 443–467.

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior research methods*, *40*, 183–190.

Vozalis, E., & Margaritis, K. (2003). Analysis of recommender systems algorithms. In *Proceedings of the 6th Hellenic European Conference on Computer Mathematics and its Applications (HERCMA-2003)*. Athens, Greece.

Wang, S., & Bond, F. (2013). Building the chinese open wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources* (pp. 10–18). Asian Federation of Natural Language Processing.

Ward Church, K., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*.

Warker, J. A., & Dell, G. S. (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 387–398.

Warrington, E. K. (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology*, *27*, 635–657.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*, 829–853.

Wertheimer, M. (1923). Undersuchungen zur lehre von der gestalt. *Psychologische Forschung*, *4*, 301–350. Reprinted in WertheimerM1938L.

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*, 1272–1288.

Williams, J. N. (1994). The relationship between word meanings in the first and second language: Evidence for a common, but restricted, semantic code. *European Journal of Cognitive Psychology*, *6*, 195–220.

Williams, J. N. (1996). Is Automatic Priming Semantic? *European Journal of Cognitive Psychology*, *8*, 113–162.

Williams, J. N. (2005). Learning Without Awareness. *Studies in Second Language Acquisition*, *27*, 269–304.

Williams, J. N. (2009). Implicit learning in Second Language Acquisition. In W. C. Ritchie, & T. K. Bhatia (Eds.), *The New Handbook of Second Language Acquisition* (pp. 319–353). Emerald Group Publishing Limited.

Wittgenstein, L. (1922). *Tractatus logico-philosophicus*. London: Routledge & Kegan Paul. Ogden, Charles Kay, Trans.

World Health Organization (2016). *World Health Statistics*. Technical Report World Health Organization.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272.

Xu, R., Gao, Z., Pan, Y., Qu, Y., & Huang, Z. (2008). An integrated approach for automatic construction of bilingual chinese-english wordnet. In J. Domingue, & C. Anutariya (Eds.), *The Semantic Web: 3rd Asian Semantic Web Conference* (pp. 302–314). Springer volume 5367.

Xue, N. (2003). Chinese word segmentation as character tagging. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing* (pp. 29–48).

Yao, B., Vasiljevic, M., Weick, M., Sereno, M. E., O'Donnell, P. J., & Sereno, S. C. (2013). Semantic size of abstract concepts: It gets emotional when you can't see it. *PLoS ONE*, *8*, e75000.

Yap, M. J., Liow, S. J. R., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, *42*, 992–1003.

Yates, M. (2005). Phonological Neighbors Speed Visual Word Processing: Evidence From Multiple Tasks. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 1385–1397.

Yates, M., Locker, L., & Simpson, G. B. (2003). Semantic and phonological influences on the processing of words and pseudohomophones. *Memory & Cognition*, *31*, 856–866.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, *37*, 891–921.

Zeiler, M. D. (2012). ADADELTA: An adaptive learning rate method. *CoRR*, *abs/1212.5701v1*.

Zhang, H. (2007). Numeral classifiers in mandarin chinese. *Journal of East Asian Linguistics*, *16*, 43–59.

Zhang, S., & Schmitt, B. H. (1998). Language-dependent classification: The mental representation of classifiers in cognition, memory, and ad evaluations. *Journal of Experimental Psychology: Applied*, *4*, 375–385.

Zipf, G. K. (1935). *The Psycho-Biology of Language*. Boston, MA: Houghton Mifflin Co.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Oxford, UK: Addison-Weasley Press.

Ziv, J., & Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, *24*, 530–536.

Zubin, D. A., & Köpcke, K.-M. (1984). Affect classification in the german gender system. *Lingua*, *63*, 41–96.

Zwitserlood, P., & Schriefers, H. (1995). Effects of sensory information and processing time in spoken-word recognition. *Language and Cognitive Processes*, *10*, 121–136.

Appendix A

Details for the in-text studies

This Appendix describes the statistical tests reported throughout the thesis without being part of a specific study. The parentheses in the section titles point to the page where the test appears.

A.1 Generating experimental stimuli from the FSG (p. 4)

Using the finite state grammar in Fig. 1.1 we were able to generate 200 unique strings of size between 6 and 8 characters. We then used these strings to calculate n-gram probabilities; because of the very low probability of some strings we log-transform these probabilities and then add them in log space (instead of multiplying them). In order to derive 'ungrammatical' strings from the same alphabet we shuffle grammatical strings and select those which cannot be parsed by the FSG. We then calculate the probability of each ungrammatical string by again adding the log probabilities of its constituent n-grams (derived from the grammatical strings). In the case where a certain bigram or trigram could not be found in the transition matrix (for example, there is no P → V transition in the grammar), we impute the probability by assigning the lowest possible probability in the grammatical corpus. An independent samples t-test revealed a significant difference between the two conditions $t(306.33) = 18.208$, $p < 0.001$.

For the GGJ2006 algorithm we used a vanilla implementation obtained from the author's website.[1] We kept most of the parameters at their default values except for the number of iterations (5000) and the prior probability of a boundary (0.3) as it was found that these parameters yielded higher probability for the corpus. We used the output frequencies as our chunk probabilities as above ensuring that each character was also a 'chunk' (in case where it wasn't we used the least probable chunk). In order to derive probabilities for the

[1] http://homepages.inf.ed.ac.uk/sgwater/software/dpseg-1.2.1.tar.gz

Table A.1 Transitional probabilities from the FSG depicted in Fig. 1.1

LHS		RHS	P(RHS|LHS)
START	→	T	0.5
START	→	V	0.5
V	→	X	0.25
V	→	V	0.25
V	→	P	0.25
V	→	S	0.25
P	→	S	0.25
P	→	X	0.25
T	→	P	0.25
T	→	T	0.25
P	→	P	0.25
P	→	T	0.25
T	→	S	0.25
T	→	X	0.25
X	→	X	0.5
X	→	V	0.5
S	→	END	1.0

ungrammatical strings, we iteratively chunked each string using the chunks from most frequent to least frequent (e.g., 'VTTSPX' → 'V', 'TTS', 'PX').

A.2 Generating the data for Fig. 1.3 (p. 16)

The data in Fig. 1.3 were generated by randomly sampling datapoints $\mathbf{x} \in \mathbb{R}^2$ from a uniform distribution ranging from $[-1, 1]$. The distance of each randomly sampled vector from the centre $\mathbf{0} = (0, 0)$ determined the class of the datapoint. Datapoints in Class A were selected such that $\left[d(\mathbf{x}, \mathbf{0}) < \frac{1}{3}\right]$ while for Class B $\left[\frac{2}{3} < d(\mathbf{x}, \mathbf{0}) < 1\right]$ where the function d is the Euclidean distance between the vector $\mathbf{x}$ and the centre $\mathbf{0}$. A neural classifier as shown in Fig. A.1 was used to fit the data. We used a tanh activation function for the hidden layer and a sigmoid for the output. The objective of the network was to minimize the binary crossentropy between the predicted and the actual class optimized with `adadelta` (Zeiler, 2012). After 500 epochs of training the crossentropy error was 0.0017. In order to derive the hidden layer representations we fed the same input patterns to the network (without updating the weights this time) recording the activity of the hidden layer.

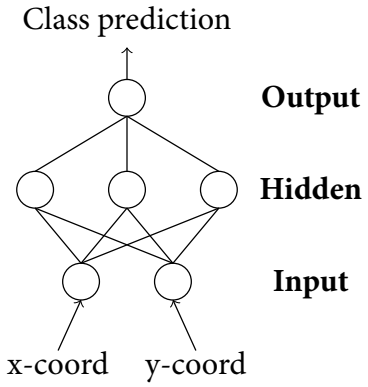


Figure A.1 Feedforward neural network used to project the two-dimensional input in Fig. 1.3a in three-dimensional space as shown in Fig. 1.3b. The activation function for the hidden layer was tanh and a sigmoid for the output. The network was trained for 500 epochs using binary crossentropy as the cost function and optimizing with `adadelta`. The error after training was 0.0017.

A.3 Clockwork orange (p. 35)

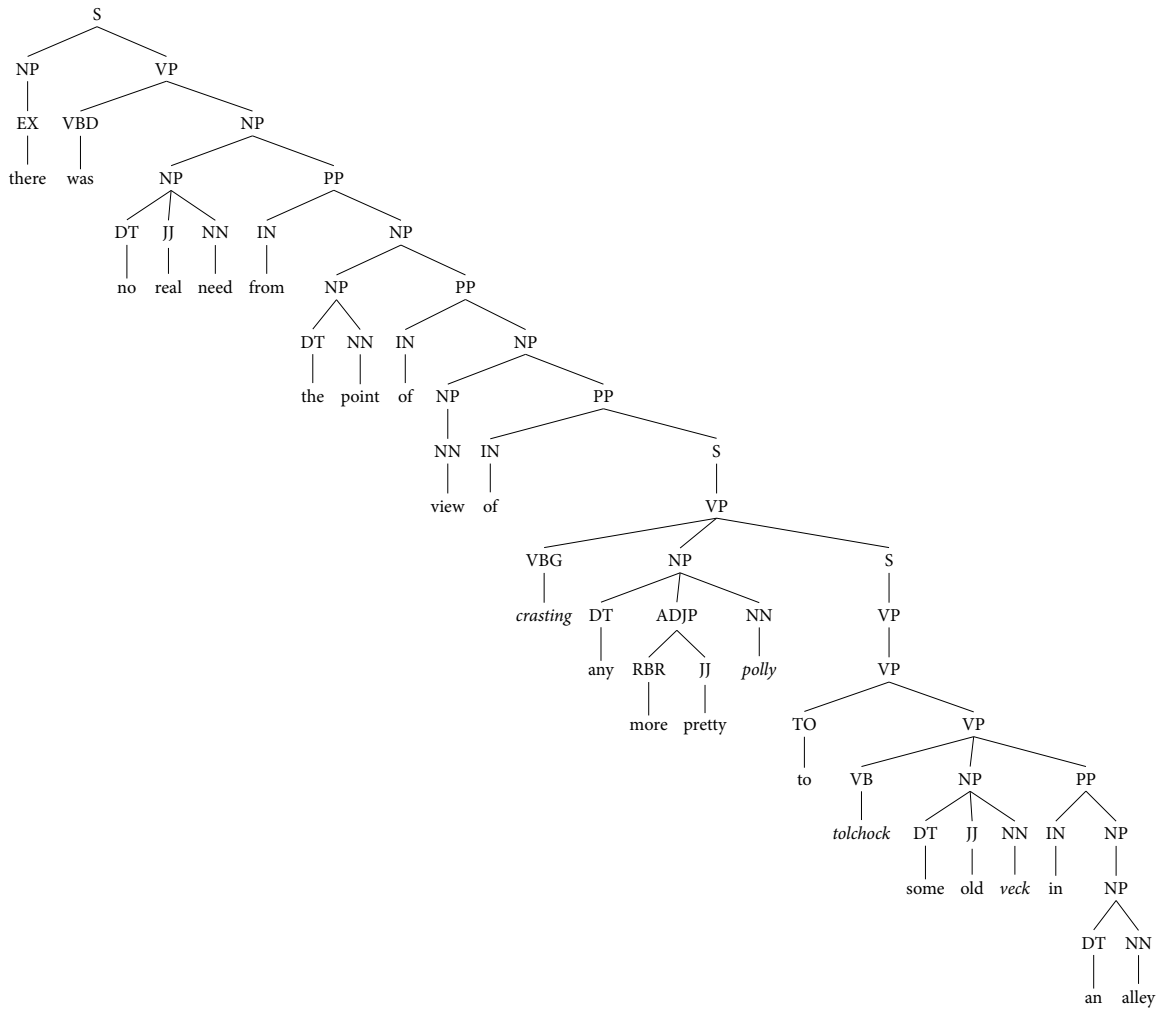


Figure A.2 Partial syntactic tree for the *Clockwork Orange* quotation in the beginning of Chapter 2 derived using the Stanford Parser. The Stanford Parser is a statistical parser that uses probabilistic knowledge gained from oracle-parsed sentences. Despite the unknown words (in *italics*) the parser is able to derive the structure of the sentences using statistical knowledge.

A.4 Semantic Priming Project

Example (p. 59)

The figures were obtained by subsetting the original dataset so that we retain trials on the 200ms SOA condition, discard reaction times below 200ms (two datapoints) and above 2500ms (two datapoints), and keep only the correct responses (two datapoints) (in all we discard in 6% of the data). While comparing only two word pairs is not very meaningful, in the remaining trials the difference between the related and unrelated condition was significant $t(35.62) = 2.09$, $p = 0.04$ (with Welch's correction).

Preprocessing (p. 69)

We outline here the preprocessing steps for the SPP dataset. Firstly, all the inconsistencies of the original dataset in the relation names were normalised (e.g., 'antonymn' → 'antonym', empty strings to 'unclassified'). Subsequently, we perform a stratified three-way split with respect to the relation between the prime and the target. In other words, we need to ensure that each split of the dataset contains a roughly proportional amount of each relation (e.g., the training set contains 64% of the antonymic relations). To do so, for each relation, we keep 80% words and add them to the *training* set and retain the rest for the *testing* set. We then repeat the same process on the *training* set to obtain the final *training* and *validation* sets. That is, the *validation* set is 20% of the *training* set from the first split, but 16% overall. This process resulted in three sets (training: 4242 word pairs, testing: 1335 word pairs and validation: 1067 word pairs). None of the predictor variables outlined in §3.3.2 varied significantly between sets in a one-way ANOVA (all $Fs < 1$), ensuring that there are no biases between splits.

Since *lasso* cannot handle matrices with missing values, we *impute* the mean of the column (i.e., the covariate) to any empty cells. We then scale each predictor to be centred around zero by subtracting the column mean from each value and then dividing by the standard deviation. We then perform feature selection using *lasso* with 10-fold cross-validation. That is, we split the dataset into ten parts, iteratively training the model on the nine and testing on the tenth. We set a `tolerance` parameter at 10^{-4} such that if the updates are lower than this value, the training should stop early. Table A.2 shows the results of an ordinary least squares regression using the variables selected from the *lasso* procedure as predictors and the z-transformed reaction times (NT200MS.Z) to each word pair as the dependent variable. All the predictors were significant in explaining portions of the variance.

Table A.2 Summary of the baseline model used in §3.4. Ordinary Least Squares Regression results using the variables selected from the *lasso* procedure as predictors and the z-transformed reaction times (NT200MS.Z) to each word pair as the dependent variable.

Predictor	β	SE	t	$P > \|t\|$	[0.025	0.975]
Prime						
Word Length	0.0178	0.003	5.570	0.000	0.012	0.024
Word Frequency	−0.0157	0.003	−4.870	0.000	−0.022	−0.009
Target						
Word Length	0.1362	0.004	33.250	0.000	0.128	0.144
Word Frequency	−0.0679	0.003	−20.748	0.000	−0.074	−0.062
Orthographic Neighbours	0.0127	0.004	3.203	0.001	0.005	0.021

Dep. Variable:	NT200MS.Z	**R-squared:**	0.324
Method:	Least Squares	**Adj. R-squared:**	0.323
No. Observations:	6644	**F-statistic:**	636.0
Df Residuals:	6639	**Prob (F-statistic):**	0.00
Df Model:	5	**Log-Likelihood:**	−108.54
AIC:	227.1	**BIC:**	261.1

Note: Prime and *Target* stand for variables related to the prime or the target word, respectively.

The preprocessed SPP dataset together with the IDS of the sets used are made publicly available for future model comparison.[2]

A.5 Evaluating dimensionality reduction in WordNet (p. 131)

We find the optimal dimensionality for the WordNet vectorised representations using the two similarity datasets (Miller & Charles, 1991; Rubenstein & Goodenough, 1965) outlined in §3.6. The two datasets combined contain 96 word pairs, and the best graph distance metric yields an average Pearson correlation on the two datasets of ≈ 0.82. Following Budanitsky & Hirst (2006), for each pair, we choose the two synsets that maximise the Leacock & Chodorow (1998) similarity metric. Subsequently, we transform these synsets into their vectorised representations and then perform dimensionality reduction using SVD on various dimensions. We chose to explore vectors of size $\{10, 50, 100, 200, 300, 500, 1000\}$ as well as a non-reduced baseline. We plot the results in Fig. A.3; we see that increasing the dimensionality improves the Pearson product-moment correlation coefficient until it plateaus around 1000 dimensions. The

[2] https://da352.user.srcf.net/datasets/spp.csv

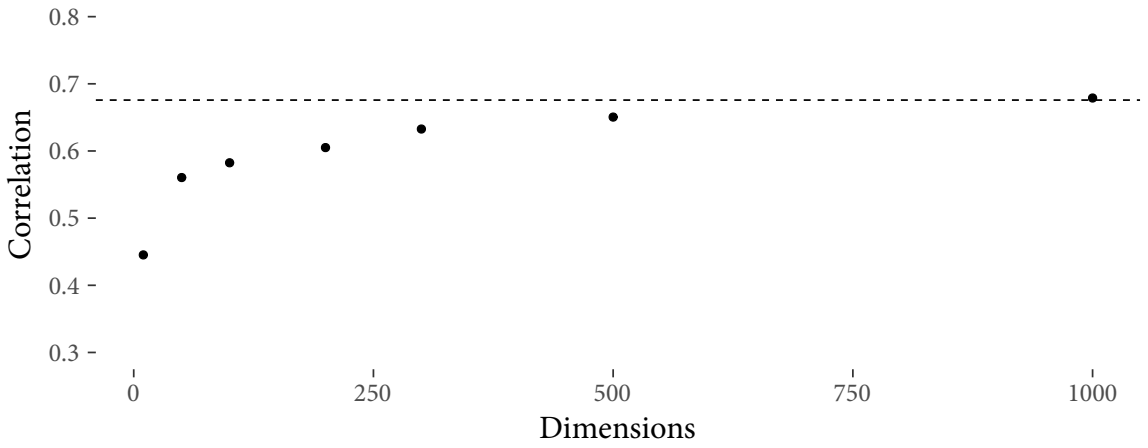


Figure A.3 Comparison of the correlation coefficients between WordNet vectorised representations and human similarity ratings (see text). The x-axis shows the increasing number of dimensions achieved through SVD while the y-axis plots the Pearson product-moment correlation coefficient. The dashed line indicates the non-reduced baseline.

reduced matrix achieves a marginally higher correlation than the non-reduced one warranting our use of the reduced matrix. However, we note that our results are lower than the average correlation achieved with the graph distance metrics.

A.6 Training Italian neural embeddings (p. 161)

We obtain neural embeddings for the Italian words using the itWaCs corpus (≈ 1 billion words). Since we cannot follow a procedure similar to the other languages (Chinese, Dutch, English, French, and Malay) where we train the neural embeddings to match the LDRT datasets, we set the model parameters empirically based on the parameter sets in Table 4.5. More specifically, we use the same parameters as we do for the French embeddings setting the dimensionality of each vector to 600 units, the window size to 18, and both the subsampling threshold and the minimum occurrences of each word to zero.

A.7 Phonological correlates of semantics (p. 175)

We examine whether it is possible to classify semantic variables (e.g., *animacy*) from phonological features. We extract phonological features for two semantic categories obtaining features from words in WordNet which fall in any of the experimental semantic categories (see Table A.3). From these sets, we excluded the multiword expressions (e.g. `beast_of_burden`) along with words that did not exist in the Carnegie Mellon's pronouncing dictionary (see below). Additionally, in WordNet, a word can appear in more than one of the semantic categories we are interested. For example, *oxygen* can follow the paths `oxygen > element > substance > matter > **physical_entity** > entity` and `oxygen > element > substance > part > relation > **abstraction** > entity`. We eliminate such inconsistencies by (1) following the first (i.e., most frequent) path and (2) ensuring that each word exists only once in the dataset either as *category A* or *category B*.[3] We then pick a random set of 1000 words for each category, ensuring that we do not include words found in any of the experimental lists. We then transform the words into their respective phonological representations and use them as input to a classifier (see below).

We train a *classifier* to learn to distinguish between semantic groups (i.e., animate / inanimate) then test its classification performance both on a novel test set and on the stimuli used in the behavioural experiments outlined in §1.5. In our simulations, we used both logistic regression and a Support Vector Machine (with a linear kernel). However, since the Support

[3] For example, concrete = $\texttt{physical\_entity.n.01} \cap \texttt{abstraction.n.01}^c$

Vector Classifier consistently outperformed simple logistic regression, we only report these results here.

A Support Vector Classifier tries to find a hyperplane in which the training data would be linearly separable. Concretely, given a matrix of training vectors $\mathbf{x} \in \mathbb{R}^p$ and a vector of labels $y \in \{1, -1\}^n$, the SVC seeks to solve the following problem,

$$\min_{\theta} C \sum_{i=1}^{n} [y_i(\mathbf{w}^\top \phi(x_i) + b)] \frac{1}{2} \sum_{i=1}^{n} \theta_i^2 \tag{A.1}$$

where n is the number of training examples (here 1000), p the dimensionality of the input (here 15 for the phonological features). C is a regularization parameter which controls the trade-off between margin and classification (in the experiment this was set to 1). The weights w are learnable coefficients and we use those to select the most important features.

We split the dataset in three parts; (1) the *training set* (64%) which is used to train the Support Vector Classifier, (2) the *testing set* (20%) which is used to assess the performance (see below) and (3) the *validation set* (16%) which is reserved for parameter selection. The evaluation of each model was done by keeping track of the F_1 score (Eq. A.4), which is the harmonic mean of the *precision* (i.e., how many selected items were relevant) and the *recall* (i.e., how many relevant items were selected), on the validation set:

$$\text{Recall} = \frac{\text{True positives}}{\text{False positives} + \text{True positives}} \tag{A.2}$$

$$\text{Precision} = \frac{\text{True positives}}{\text{False negatives} + \text{True positives}} \tag{A.3}$$

$$F_1 \text{ score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{A.4}$$

We use *lasso* classification for both our predictions and the feature selection procedure (described in detail in §3.3.1). *Lasso* forces the magnitude of each weight vector to be as small as

Table A.3 WordNet synsets used in the simulations

Semantic Category	WordNet synsets containing
Animate	`living_thing.n.01`
Inanimate	`artifact.n.01`
Abstract	`abstraction.n.06`
Concrete	`physical_entity.n.01`

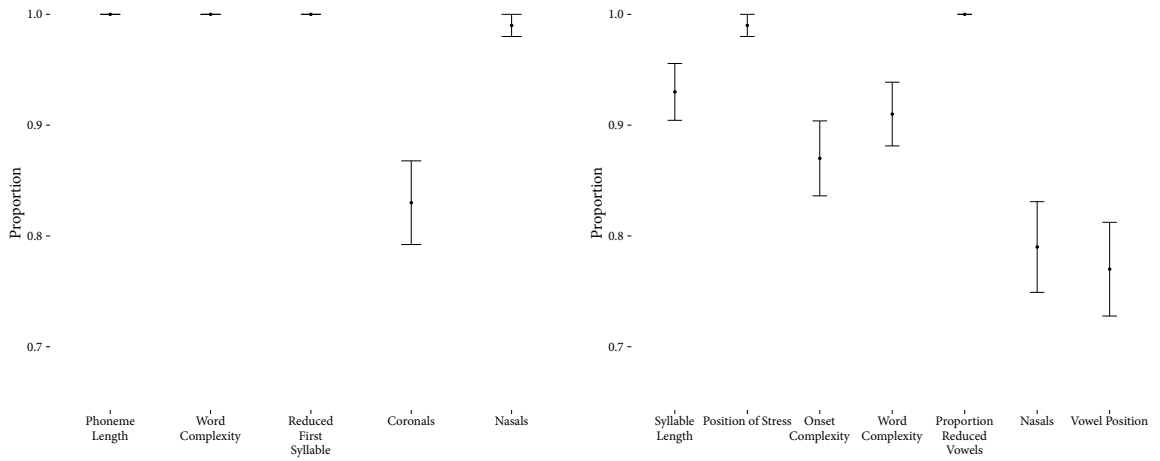
Table A.4 Performance of the trained classifier on the datasets used in the behavioural experiments. Standard error of the mean in parentheses.

	Animacy	Concreteness
F_1 score	0.51 (0.001)	0.62 (0.004)
Precision	0.42 (0.002)	0.54 (0.002)
Recall	0.66 (0.001)	0.76 (0.01)

possible. The weights of irrelevant features, therefore, will be close to zero whereas important features will have values other than zero. We carry our feature selection process, then, as follows; looking solely at the validation set, we eliminate features based on whether (1) their θs are larger than some threshold value τ and (2) F_1 scores improve on this set. Starting from $\tau = 0.05$ and moving in steps of 0.05 we eliminate features for which $|\theta_j| < \tau$. This procedure carries on until the scores on the validation set have not improved for 3 steps, which is when we stop the process and use the remaining features to train the classifier reporting the scores of the unseen testing set. This guarantees a more unbiased computation of the testing scores. We repeat the same process 100 times and report the averages for the F_1 scores and the times each feature survived the elimination process. The evaluation scores for each semantic distinction are shown in Fig. A.5 and the most relevant features for each category are shown in Fig. A.4.

Regarding the feature elimination procedure, a Mann-Whitney U Test did not reveal any significant differences between the full and the reduced featuresets for either the animate/inanimate distinction ($U = 5657.5, Z = 1.6065, n.s.$) or the abstract/concrete one ($U = 4507.5, Z = -1.2034, n.s.$), which points to the direction that the eliminated features were adding noise not contributing to the classification accuracy. All the statistics reported below are on the reduced featureset. F_1 scores were significantly higher for the concreteness distinction than for animacy ($U = 8765, Z = -9.1994, p = 0$). We also ran a linear regression model with `semantic.distinction` and `metric` as the predictors and the `score` as the dependent variable. This permits us to explore whether there are any interactions between the different metrics. Interestingly, we found a `semantic.distinction` × `metric` interaction ($\beta = -0.06, t = -25.00, p = 0$) as shown in Fig. A.5c.

The interpretation of these metrics aids at understanding the above interaction. Eqs. (A.2-A.3) show what they are really measuring. Consider that the true labels are $[0, 0, 0, 0, 1, 1, 1, 1]$ (where 0 could be *abstract* and 1 *concrete*). If the vector of predictions of the model was $[1, 1, 1, 1, 1, 1, 1, 1]$ then the recall would be 1 as the model selected all the relevant items. However, its precision would be 0.5 as only half of its predictions were relevant. In this extreme



(a) Animate – Inanimate **(b)** Abstract – Concrete

Figure A.4 Phonological features predictive of semantic classes. Proportion of how many times (out of 100) each feature was considered predictive of the semantic class. Since each of the 100 simulations was done on a different splitting of the dataset some features might not have been predictive in a specific split. Here we show the features which survived the above procedure more than 75% of the time. Error bars denote the standard error of the mean.

case, the harmonic mean of the two scores (i.e., the F_1 score) would be 0.66 (Fig. A.5d). According to this explanation then, the *precision* provides a lower bound of the model's performance. With this in mind, we see that while the F_1 scores are significantly higher for the *concreteness* distinction, suggesting that the model is more able to predict this from phonological features, the *precision* scores are significantly higher for the *animacy* distinction ($U = 1043.5$, $Z = 9.6673$, $p = 0$) suggesting that the model is better able to distinguish animate from inanimate concepts. This is further supported by the fact that the features predictive of *animacy* have lower variance than those of *concreteness* on different subsets of the same dataset. We interpret this as the model finding a consistent solution to the predict *animacy* whereas in the case of *concreteness* the model kind of 'overfits' to the specific subset of the data.

In sum, there are two main findings from this exploratory study; firstly, when a sophisticated machine learning model is trained to associate phonological representations with meaning, then it can mildly predict the semantic class of a novel word. These results agree with previous results from the literature (e.g., Monaghan et al., 2014a) in that there are mild regularities in the phonological representations of words. What is more, in this study we attempted a coarse exploration without looking into more specific subgroupings (e.g., liquids

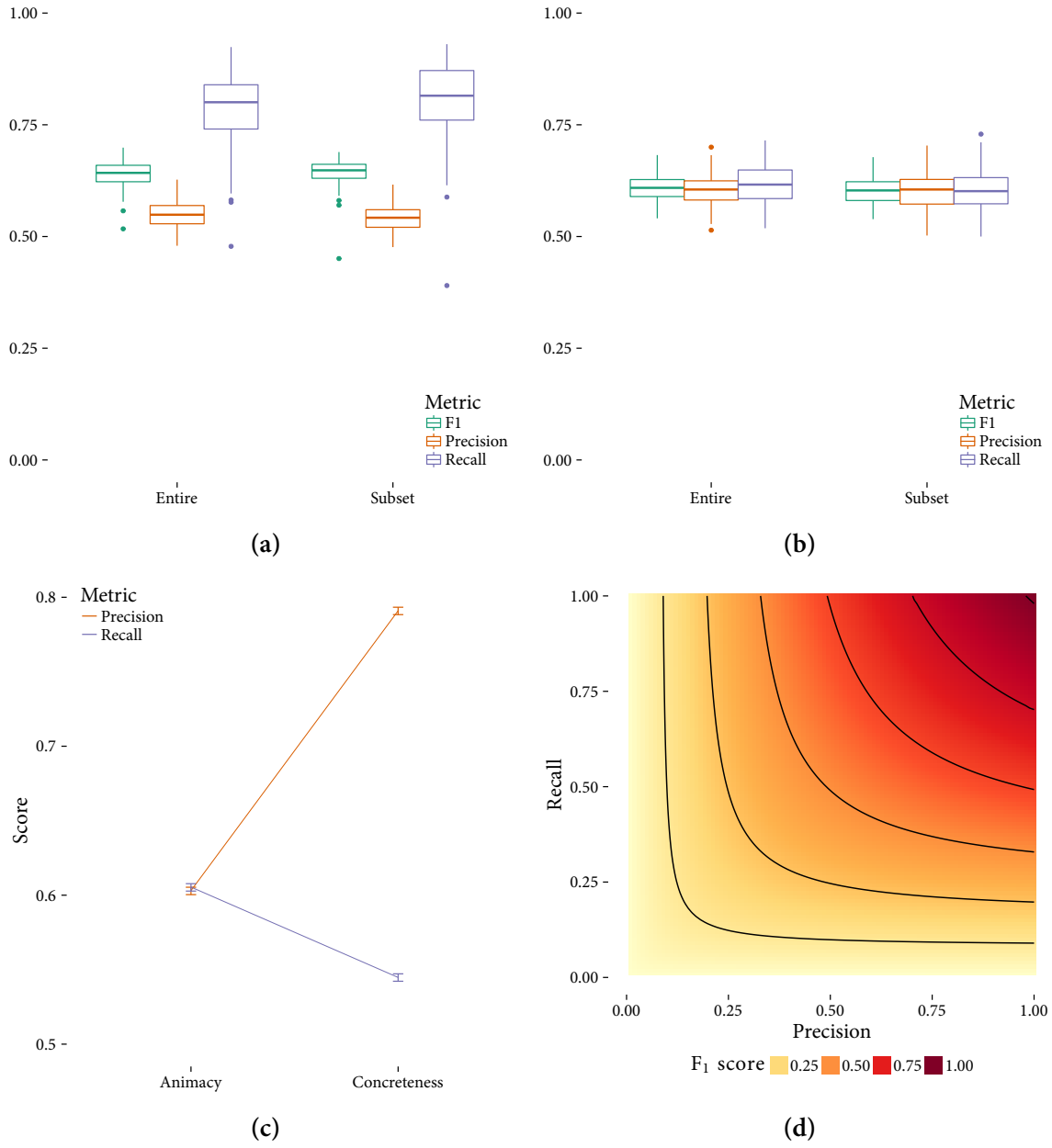


Figure A.5 Results of the semantic classification according to phonological features. (a) Boxplots of the three metrics used on both the full featureset and the reduced one for the abstract/concrete distinction. (b) Similar boxplots for the animate/inanimate distinction. (c) By semantic distinction interaction of the recall and precision metrics. (d) F_1 scores as a function of precision and recall.

or chemical elements). Secondly, Table A.4 shows that even if participants attempted to solve the SIL tasks solely on phonological grounds, then their lower bound performance would still be very close to chance (albeit statistically higher).

Appendix B

Simulation Details

B.1 Initialisation of the weights

We initialise the learnable matrices using the scheme proposed by Glorot & Bengio (2010) according to which each matrix $\mathbf{W}$ is sampled from a uniform distribution U as follows:

$$\mathbf{W} \sim U\left[-\frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}}, \frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}}\right] \tag{B.1}$$

where n_j is the size of the jth layer of the network. The above initialisation method leads to substantially faster convergence and less saturation in the backpropagation of the gradients.

B.2 Bayesian Optimiser parameter spaces

Table B.1 offers the parameter spaces for all the models tested in §§ 3.4 and 4.5 (no new simulations were performed for §§ 3.5 and 4.4). The 'Simple' column shows the values used in the simulations in readable format. However, since the Bayesian Optimiser can only accept either ordered sequences or unordered sets, for transparency, we also include the values used to perform the simulations (note that the 'Simple' and 'Bayesian Optimiser' columns are equivalent). Concretely, we construct sequences as noted in the intervals on the right hand side, and then compute the actual parameter value programmatically as seen on the left. To illustrate why this is needed, consider the set $\{256, 512, 1024, 2048\}$; if we had passed these values as a list, then the optimiser would not be able to know that 512 is twice 256. On the other hand, we could have passed a range 256 – 2048, which, however, would have the unfortunate effect of exploring all possible values between 256 and 2048. To preserve the ordinal sequence of the values we then pass a sequence of consecutive numbers (which the optimiser is allowed

Table B.1 Parameter spaces for each of the models tested in §§ 3.4 and 4.5. An explanation for each parameter can be found in the description of each model Chapter 2 (and the references therein). 'Simple' presents the ranges of each parameter in a readable format, whereas 'Bayesian Optimiser' shows the sequence that was sent to the optimiser along with the computation that was performed.

Model	Possible Values	
	Simple	Bayesian Optimiser
BEAGLE		
Dimensionality[1]	$[256, 512, \ldots, 2048]$	$2^i, \forall i \in [8, 12]$
Semantic Type	{Context, Ordering, Composite}	
COALS		
Dimensionality	14000	
Reduce[2]	True, False	
Reduced Dimension	$[100, 200, \ldots, 400]$	$100i, \ \forall i \in [1, 4]$
HAL		
Window Size	$[3, 5, 7]$	$2i + 1, \ \forall i \in [1, 3]$
Retain	14000	
LSA		
Dimensionality	$[100, 200, \ldots, 700]$	$100i, \ \forall i \in [1, 7]$
Neural Embeddings[3]		
Hierarchical Softmax	True, False	
CBOW mean	True, False	
Negative samples	$[2, 4, \ldots, 20]$	$2i, \ \forall i \in [0, 10]$
Subsampling threshold	$\{0\} \cup [10^{-1}, \ldots, 10^{-5}]$	$\{0\} \cup \{10^{-i}, \ \forall i \in [1, 5]\}$
Window Size	$[2, 4, \ldots, 20]$	$2i, \ \forall i \in [1, 10]$
Minimum occurrences	$[0, 5, \ldots, 30]$	$5i, \ \forall i \in [0, 6]$
Skip-gram	True, False	
Dimensionality	$[50, 100, \ldots, 600]$	$50i, \ \forall i \in [1, 12]$
Random Indexing		
Dimensionality	$[256, 512, \ldots, 2048]$	$2^i, \ \forall i \in [8, 12]$
Window Size	$[3, 5, 7]$	$2i + 1, \ \forall i \in [1, 3]$

[1] We use powers of two for the size of the vectors since we compute the circular convolution summation using Fast-Fourier Transformations which are more efficient when the dimensionality is a power of two.

[2] Whether or not the dimensionality of the initial matrix is reduced using SVD.

[3] An explanation for the parameters that are not included in §2.3.1 can be found in the original paper (Mikolov et al., 2013b).

to choose from) and then compute the value that the DSM is trained on.[1] Furthermore, some parameter values are mutually exclusive; if, for example, no dimensionality reduction was proposed in COALS, then the 'Reduced Dimensionality' parameter was redundant. In these cases, the program ignored the redundant variables and either continued with the simulation or proposed a new set of parameters (if a value had already been found).

B.3 Computing the semantic neighbourhoods

We calculate the semantic neighbourhood density by firstly substituting $\mathbf{M}$ (i.e., the representation matrix) with the $L_{2,1}$ normed matrix $\|\mathbf{M}\|_{2,1} \triangleq \sum_{j=1}^{|\mathcal{V}|} \left(\sum_{i=1}^{D} |M_{ij}|^2\right)^{1/2}$, where $|\mathcal{V}|$ is the size of the vocabulary, and D the dimensionality of the word vectors. Normalising the matrix this way enables us to compute the cosine similarity of a single word to every other word in the vocabulary by simply taking the dot product between the word vector and the representation matrix ($\mathbf{M} \cdot \mathbf{w}^\top$). We speed up computation by feeding batches of word vectors instead of a single vector and subsequently concatenating the results. Concretely, we take $\mathbf{M} \cdot \mathbf{W}_i$, where $\mathbf{W}_i \in \mathbb{R}^{B \times D}$ is the i-th batch and B is the size of the batch (here set to 2000). The above results in a $D \times |\mathcal{V}|$ matrix, which contains the similarity of every word in the batch to every other word in the vocabulary. For every row in this matrix, we select the top 10 values (excluding the highest value, as this should be the similarity of the word with itself) and average them together. We repeat the same procedure for all the batches and then combine the resulting vectors.

For the HAL and COALS simulations we forced the models to retain the 10^5 most common words. We do so in order to reduce the size of the resulting co-occurrence matrix as the size of each word-vector is substantially higher than the rest of the models. For example, retaining the full vocabulary would require ≈36GB of memory space (assuming double precision floats) or ≈18GB (assuming single precision). Considering that this matrix would need to be retained in memory while we perform the above computations for the semantic neighbourhood density, the computational demands can quickly become problematic. Instead, retaining the top 10^5 words reduces the total size of the matrices by a third (≈5GB –assuming single precision) which easily fit in the memory of a modern computer.

[1] Actually, regarding the `dimensionality` values for BEAGLE and Random Indexing, we pass the sequence $[0, 1, 2, 3, 4]$ and then compute $2^{(i+8)}$, $\forall i \in [0, \ldots, 4]$. However, for simplicity we show the values directly.

Table B.2 Vowel height and vowel position values used to generate the phonological representations. The values were derived by splitting the two-dimensional vowel space in steps of 0.25 using 0 to 3 as the minimum/maximum for vowel height and 0 to 2 for vowel position (e.g., a *close* vowel would get a height value equal to 0, whereas an *open-mid* would be 2).

ARPABET	IPA	Vowel height	Vowel position
AA	ɑ	3	2
AE	æ	2.5	0
AH	ʌ	2	2
AO	ɔ	2	2
AW	aʊ	1.75	1.75
AX	ə	1.5	1
AY[1]	aɪ	1.75	0.25
EH	ɛ	2	0
ER	ɚ	2	1
EY	eɪ	0.75	0.25
IH	ɪ	0	0.5
IY	i	0	0
OW	oʊ	0.75	1.75
OY	ɔɪ	1.25	1.25
UH	ʊ	0	1.5
UW	u	0	2

Note: The values for the feature Stressed Vowel Position were the same as 'Vowel Position'.
[1] The values for the diphthongs (see also, Table B.4) were derived by averaging the values of each vowel.

B.4 Phonological Cues

We derive phonological feature vectors using the scheme outlined in Monaghan et al. (2005) (see also, §6.3.2). A brief explanation of each feature along with some descriptive statistics is given in Table B.3. Table B.2 shows the values used to chunk the two-dimensional vowel space. Finally, Table B.4 is a supplement to Table B.3 showing the distribution of phonemes in each category.

Table B.3 Description of the features used in §6.3. For more details, see Monaghan et al. (2005). The reported statistics were computed on 4000 words marked as one of abstract, concrete, animate or inanimate.

Feature	Mean	SD	Min	Max
Word level				
Length in phonemes	6.17	(2.26)	2.00	15.00
Length in syllables	1.90	(0.94)	1.00	6.00
Presence of stress	1.00	(0.02)	0.00	1.00
Position of stress	1.33	(0.63)	0.00	3.00
Syllable level				
Onset complexity	1.05	(0.59)	0.00	3.00
Word complexity	0.61	(0.09)	0.25	0.83
Proportion of reduced vowels	0.04	(0.19)	0.00	1.00
Reduced first syllable	0.41	(0.49)	0.00	1.00
Phoneme level				
Coronals	0.59	(0.26)	0.00	1.00
Initial /ð/	0.00	(0.00)	0.00	0.00
Final voicing	2.58	(0.65)	0.00	2.00
Nasals	0.11	(0.12)	0.00	0.67
Stressed vowel position	0.82	(0.85)	0.00	2.00
Vowel position	1.04	(0.59)	0.00	2.00
Vowel height	1.49	(0.70)	0.00	3.00

Table B.4 Division of phonemes in the ARPABET format

Consonants						Vowels		
Semivowels	Stops	Affricates	Nasals	Fricatives	Liquids	Diphthongs	Monophthongs	R-colored
Y (j)	P (p)	CH (ʧ)	M (m)	F (f)	L (l)	EY (eɪ)	AO (ɔ)	ER (ɝ)
W (w)	B (b)	JH (ʤ)	EM (m̩)	V (v)	EL (l̩)	AY (aɪ)	AA (ɑ)	AXR (ɚ)
Q (ʔ)	T (t)		N (n)	TH (θ)	R (r)	OW (oʊ)	IY (i)	EH R (ɛr)
	D (d)		EN (n̩)	DH (ð)	DX (ɾ)	AW (aʊ)	UW (u)	UH R (ʊr)
	K (k)		NG (ŋ)	S (s)	NX (ɾ̃)	OY (ɔɪ)	EH (ɛ)	AO R (ɔr)
	G (g)		ENG (ŋ̩)	Z (z)			IH (ɪ)	AA R (ɑr)
				SH (ʃ)			UH (ʊ)	IH R (ɪr)
				ZH (ʒ)			AH (ʌ)	AW R (aʊr)
				HH (h)			AX (ə)	
							AE (æ)	

Appendix C

Dataset Details

Here we detail the datasets used in the experiments. For each simulation, we introduce to the model the index of the word vector (see Appendix B) from the embedding matrix corresponding to the relevant noun. For each of the determiners, we use one-hot encoding. That is, we assign a unique vector to each determiner where only one element is set to 1 and the rest to 0.

Each section contains the training and testing stimuli used in the corresponding simulation. Each table containing training stimuli has four columns describing the noun used, the relevant semantic property of that noun (e.g., animate), the novel word with which it was paired, and the meaning of that novel word. Here we report the determiners from the behavioural experiments for illustration purposes. The model never sees the information contained in the semantic variable or the meaning columns. Each testing table contains similar information with the addition of the incorrect alternative for that specific noun. For example, if during training the participants saw 'gi dog' the ungrammatical alternative would be 'ro dog' (correct meaning, incorrect semantic property).

C.1 Animate / Inanimate (p. 133)

The stimuli in these experiments come from either Williams (2005, Experiment 1) (for the English simulations) or Chen et al. (2011, Experiment 1) (for the Mandarin Chinese). Chen et al. (2011) provide the actual Chinese characters used in the experiments, so we did not use the MDBG to translate the English words.

Table C.1 Training materials for the 'Animate / Inanimate' simulations

Noun	Semantic Variable	Novel Word	Meaning
lion	animate	gi	near
bird	animate	gi	near
dog	animate	gi	near
mouse	animate	gi	near
cow	animate	gi	near
cat	animate	gi	near
snake	animate	gi	near
pig	animate	gi	near
bear	animate	gi	near
monkey	animate	ul	far
bee	animate	ul	far
dog	animate	ul	far
mouse	animate	ul	far
cow	animate	ul	far
cat	animate	ul	far
snake	animate	ul	far
pig	animate	ul	far
bear	animate	ul	far
table	inanimate	ro	near
vase	inanimate	ro	near
sofa	inanimate	ro	near
cup	inanimate	ro	near
television	inanimate	ro	near
book	inanimate	ro	near
plate	inanimate	ro	near
box	inanimate	ro	near
picture	inanimate	ro	near
stool	inanimate	ne	far
clock	inanimate	ne	far
sofa	inanimate	ne	far
cup	inanimate	ne	far

Continued on next page

Table C.1 – continued from previous page

Noun	Semantic Variable	Novel Word	Meaning
television	inanimate	ne	far
book	inanimate	ne	far
plate	inanimate	ne	far
box	inanimate	ne	far
picture	inanimate	ne	far

Table C.2 Testing materials for the 'Animate / Inanimate' simulations

Noun	Semantic variable	Correct	Incorrect	Meaning
monkey	animate	near	gi	ro
bee	animate	near	gi	ro
lion	animate	far	ul	ne
bird	animate	far	ul	ne
stool	inanimate	near	ro	gi
clock	inanimate	near	ro	gi
table	inanimate	far	ne	ul
vase	inanimate	far	ne	ul

C.2 Abstract / Concrete (p. 139)

We extract the materials for both the high- and low- similarity conditions from Paciorek & Williams (2015).

C.2.1 High similarity

Table C.3 Training materials for the 'Abstract / Concrete (High similarity)' simulations

Noun	Semantic variable	Novel Word	Meaning
force	abstract	gouble	increase
significance	abstract	gouble	increase
greatness	abstract	gouble	increase

Continued on next page

Table C.3 – continued from previous page

Noun	Semantic variable	Novel Word	Meaning
authority	abstract	gouble	increase
prestige	abstract	powter	decrease
prominence	abstract	powter	decrease
status	abstract	powter	decrease
role	abstract	powter	decrease
carbohydrates	concrete	conell	increase
oxygen	concrete	conell	increase
serotonin	concrete	conell	increase
ozone	concrete	conell	increase
nutrients	concrete	mouten	decrease
calcium	concrete	mouten	decrease
minerals	concrete	mouten	decrease
glucose	concrete	mouten	decrease

Table C.4 Testing materials for the 'Abstract / Concrete (High similarity)' simulations

Noun	Semantic variable	Correct	Incorrect	Meaning
appeal	abstract	gouble	conell	increase
impact	abstract	gouble	conell	increase
power	abstract	gouble	conell	increase
splendour	abstract	gouble	conell	increase
eminence	abstract	gouble	conell	increase
fame	abstract	gouble	conell	increase
recognition	abstract	gouble	conell	increase
strength	abstract	gouble	conell	increase
importance	abstract	powter	mouten	decrease
prosperity	abstract	powter	mouten	decrease
trust	abstract	powter	mouten	decrease
value	abstract	powter	mouten	decrease
acclaim	abstract	powter	mouten	decrease
esteem	abstract	powter	mouten	decrease
influence	abstract	powter	mouten	decrease

Continued on next page

Table C.4 – continued from previous page

Noun	Semantic variable	Correct	Incorrect	Meaning
position	abstract	powter	mouten	decrease
calories	concrete	conell	gouble	increase
histamine	concrete	conell	gouble	increase
potassium	concrete	conell	gouble	increase
sugar	concrete	conell	gouble	increase
dopamine	concrete	conell	gouble	increase
fertilizers	concrete	conell	gouble	increase
insulin	concrete	conell	gouble	increase
proteins	concrete	conell	gouble	increase
enzymes	concrete	mouten	powter	decrease
methane	concrete	mouten	powter	decrease
nitrogen	concrete	mouten	powter	decrease
vitamins	concrete	mouten	powter	decrease
aerosol	concrete	mouten	powter	decrease
glycogen	concrete	mouten	powter	decrease
hydrogen	concrete	mouten	powter	decrease
magnesium	concrete	mouten	powter	decrease

C.2.2 Low similarity

Table C.5 Training materials for the 'Abstract / Concrete (Low similarity)' simulations

Noun	Semantic variable	Novel Word	Meaning
force	abstract	gouble	increase
authority	abstract	gouble	increase
happiness	abstract	gouble	increase
value	abstract	gouble	increase
relevance	abstract	powter	decrease
prestige	abstract	powter	decrease
anger	abstract	powter	decrease
charm	abstract	powter	decrease
oxygen	concrete	conell	increase

Continued on next page

Table C.5 – continued from previous page

Noun	Semantic variable	Novel Word	Meaning
cream	concrete	conell	increase
carbon	concrete	conell	increase
cotton	concrete	conell	increase
furniture	concrete	mouten	decrease
calcium	concrete	mouten	decrease
petrol	concrete	mouten	decrease
honey	concrete	mouten	decrease

Table C.6 Testing materials for the 'Abstract / Concrete (Low similarity)' simulations

Noun	Semantic variable	Correct	Incorrect	Meaning
feeling	abstract	gouble	conell	increase
fame	abstract	gouble	conell	increase
concern	abstract	gouble	conell	increase
impact	abstract	gouble	conell	increase
surprise	abstract	gouble	conell	increase
prosperity	abstract	gouble	conell	increase
wisdom	abstract	gouble	conell	increase
understanding	abstract	gouble	conell	increase
reputation	abstract	powter	mouten	decrease
success	abstract	powter	mouten	decrease
pride	abstract	powter	mouten	decrease
anxiety	abstract	powter	mouten	decrease
quality	abstract	powter	mouten	decrease
likelihood	abstract	powter	mouten	decrease
fear	abstract	powter	mouten	decrease
esteem	abstract	powter	mouten	decrease
plastic	concrete	conell	gouble	increase
salt	concrete	conell	gouble	increase
meat	concrete	conell	gouble	increase
glass	concrete	conell	gouble	increase
sand	concrete	conell	gouble	increase

Continued on next page

Table C.6 – continued from previous page

Noun	Semantic variable	Correct	Incorrect	Meaning
paint	concrete	conell	gouble	increase
metal	concrete	conell	gouble	increase
wheat	concrete	conell	gouble	increase
chocolate	concrete	mouten	powter	decrease
wood	concrete	mouten	powter	decrease
paper	concrete	mouten	powter	decrease
glue	concrete	mouten	powter	decrease
luggage	concrete	mouten	powter	decrease
bread	concrete	mouten	powter	decrease
grass	concrete	mouten	powter	decrease
soil	concrete	mouten	powter	decrease

C.3 Perceptual features (p. 146)

The stimuli in this experiment come from Chen et al. (2011, Experiment 3). We also use the English translations of the same words to simulate the effects in English (similar to Leung & Williams, 2012)

Table C.7 Training materials for the 'Perceptual features' simulations

Noun	Semantic variable	Novel Word	Meaning
deer	big	chu	near
panda	big	chu	near
cow	big	chu	near
pig	big	chu	near
bear	big	chu	near
lion	big	chu	near
shark	big	chu	near
elephant	big	chu	near
leopard	big	chu	near
horse	big	chu	near
deer	big	guai	far
panda	big	guai	far

Continued on next page

Table C.7 – continued from previous page

Noun	Semantic variable	Novel Word	Meaning
cow	big	guai	far
pig	big	guai	far
bear	big	guai	far
lion	big	guai	far
shark	big	guai	far
elephant	big	guai	far
leopard	big	guai	far
horse	big	guai	far
cock	small	ya	near
frog	small	ya	near
monkey	small	ya	near
bee	small	ya	near
mouse	small	ya	near
cat	small	ya	near
fly	small	ya	near
insect	small	ya	near
bird	small	ya	near
tortoise	small	ya	near
cock	small	tuo	far
frog	small	tuo	far
monkey	small	tuo	far
bee	small	tuo	far
mouse	small	tuo	far
cat	small	tuo	far
fly	small	tuo	far
insect	small	tuo	far
bird	small	tuo	far
tortoise	small	tuo	far

Table C.8 Testing materials for the 'Perceptual features' simulations

Noun	Semantic variable	Correct	Incorrect	Meaning
sheep	big	chu	ya	near
kangaroo	big	chu	ya	near
tiger	big	chu	ya	near
crocodile	big	chu	ya	near
hippo	big	chu	ya	near
camel	big	chu	ya	near
donkey	big	chu	ya	near
wolf	big	chu	ya	near
sheep	big	guai	tuo	far
kangaroo	big	guai	tuo	far
tiger	big	guai	tuo	far
crocodile	big	guai	tuo	far
hippo	big	guai	tuo	far
camel	big	guai	tuo	far
donkey	big	guai	tuo	far
wolf	big	guai	tuo	far
rabbit	small	ya	chu	near
goldfish	small	ya	chu	near
snail	small	ya	chu	near
cicada	small	ya	chu	near
shrimp	small	ya	chu	near
eagle	small	ya	chu	near
ant	small	ya	chu	near
dragonfly	small	ya	chu	near
rabbit	small	tuo	guai	far
goldfish	small	tuo	guai	far
snail	small	tuo	guai	far
cicada	small	tuo	guai	far
shrimp	small	tuo	guai	far
eagle	small	tuo	guai	far
ant	small	tuo	guai	far

Continued on next page

Table C.8 – continued from previous page

Noun	Semantic variable	Correct	Incorrect	Meaning
dragonfly	small	tuo	guai	far

C.4 Language-Specific distributional cues (p. 150)

For these two experiments (Chinese and English) we use the stimuli from Leung & Williams (2014). Since the experiment on speakers of Chinese was conducted in Cantonese Chinese but our corpus was in Mandarin, we use the MDBG to get the translations of the words in Mandarin. We do so to ensure that the words used take the same classifiers in Mandarin Chinese (see main text). We also add one column which gives the English translation of the Mandarin Chinese word as given in the MDBG.

C.4.1 Mandarin Chinese

Table C.9 Training materials for the 'Language-Specific distributional cues (Mandarin Chinese)' simulations

Noun	English	Semantic variable	Novel Word	Meaning
鏈條	chain	long	gi	near
胳臂	arm	long	gi	near
香蕉	banana	long	gi	near
虹	rainbow	long	gi	near
麥	wheat	long	gi	near
薯條	french fries	long	gi	near
眉毛	eyebrow	long	gi	near
蜈蚣	centipede	long	gi	near
粉筆	chalk	long	gi	near
芹菜	celery	long	gi	near
蚯蚓	earthworm	long	gi	near
褲子	trousers	long	gi	near
輸送帶	conveyor belt	long	gi	near
香腸	sausage	long	gi	near
電線杆	electric pole	long	gi	near

Continued on next page

Table C.9 – continued from previous page

Noun	English	Semantic variable	Novel Word	Meaning
分支	branch	long	gi	near
骨頭	bone	long	gi	near
禮券	gift voucher	flat	ro	near
紙巾	paper towel	flat	ro	near
支票	check (bank)	flat	ro	near
幻燈片	slide (photography)	flat	ro	near
錫箔紙	aluminum foil	flat	ro	near
毯子	blanket	flat	ro	near
旗	banner	flat	ro	near
收藏	bookmark	flat	ro	near
卡片	card	flat	ro	near
桌布	tablecloth	flat	ro	near
地毯	carpet	flat	ro	near
證	certificate	flat	ro	near
信用卡	credit card	flat	ro	near
信封	envelope	flat	ro	near
床單	bed sheet	flat	ro	near
選票	a vote	flat	ro	near
牌照	(business) licence	flat	ro	near
毛	hair	long	ul	far
鐲子	bracelet	long	ul	far
血管	vein	long	ul	far
煙	cigarette or pipe tobacco	long	ul	far
管子	tube	long	ul	far
釣竿	fishing rod	long	ul	far
鐵軌	rail	long	ul	far
繩子	cord	long	ul	far
地洞	tunnel	long	ul	far
火把	torch	long	ul	far
雞腿	chicken leg	long	ul	far
梯子	ladder	long	ul	far
皮帶	strap	long	ul	far

Continued on next page

Table C.9 – continued from previous page

Noun	English	Semantic variable	Novel Word	Meaning
鞭子	whip	long	ul	far
紅蘿蔔	carrot	long	ul	far
尾	tail	long	ul	far
線	thread	long	ul	far
床墊	mattress	flat	ne	far
報紙	newspaper	flat	ne	far
告示牌	notice	flat	ne	far
畫	picture	flat	ne	far
紙	paper	flat	ne	far
攝	photo	flat	ne	far
明信片	postcard	flat	ne	far
海報	poster	flat	ne	far
收據	receipt	flat	ne	far
紅牌	red card (sports)	flat	ne	far
地圖	map	flat	ne	far
手帕	handkerchief	flat	ne	far
票	ticket	flat	ne	far
標籤	label	flat	ne	far
聖誕卡	Christmas card	flat	ne	far
郵票	(postage) stamp	flat	ne	far
光碟	optical disc	flat	ne	far

Table C.10 Testing materials for the 'Language-Specific distributional cues (Mandarin Chinese)' simulations

Noun	English	Semantic variable	Correct	Incorrect	Meaning
香蕉	banana	long	ul	ne	far
芹菜	celery	long	ul	ne	far
香腸	sausage	long	ul	ne	far
電線杆	electric pole	long	ul	ne	far
分支	branch	long	ul	ne	far
支票	check (bank)	flat	ne	ul	far

Continued on next page

Table C.10 – continued from previous page

Noun	English	Semantic variable	Correct	Incorrect	Meaning
毯子	blanket	flat	ne	ul	far
收藏	bookmark	flat	ne	ul	far
桌布	tablecloth	flat	ne	ul	far
信封	envelope	flat	ne	ul	far
煙	cigarette or pipe tobacco	long	gi	ro	near
管子	tube	long	gi	ro	near
繩子	cord	long	gi	ro	near
梯子	ladder	long	gi	ro	near
皮帶	strap	long	gi	ro	near
地圖	map	flat	ro	gi	near
票	ticket	flat	ro	gi	near
聖誕卡	Christmas card	flat	ro	gi	near
郵票	(postage) stamp	flat	ro	gi	near
光碟	optical disc	flat	ro	gi	near

C.4.2 English

Table C.11 Training materials for the 'Language-Specific distributional cues (English)' simulations

Noun	Semantic variable	Novel Word	Meaning
chain	long	gi	near
arm	long	gi	near
banana	long	gi	near
rainbow	long	gi	near
wheat	long	gi	near
french fries	long	gi	near
eyebrow	long	gi	near
centipede	long	gi	near
chalk	long	gi	near
celery	long	gi	near
earthworm	long	gi	near

Continued on next page

Table C.11 – continued from previous page

Noun	Semantic variable	Novel Word	Meaning
trousers	long	gi	near
conveyor belt	long	gi	near
sausage	long	gi	near
lamp post	long	gi	near
branch	long	gi	near
bone	long	gi	near
coupon	flat	ro	near
tissue	flat	ro	near
cheque	flat	ro	near
transparency	flat	ro	near
foil	flat	ro	near
blanket	flat	ro	near
banner	flat	ro	near
bookmark	flat	ro	near
card	flat	ro	near
tablecloth	flat	ro	near
carpet	flat	ro	near
certificate	flat	ro	near
credit card	flat	ro	near
envelope	flat	ro	near
bed sheet	flat	ro	near
ballot	flat	ro	near
licence	flat	ro	near
hair	long	ul	far
bracelet	long	ul	far
vein	long	ul	far
cigarette	long	ul	far
pipe	long	ul	far
rod	long	ul	far
rail	long	ul	far
rope	long	ul	far
tunnel	long	ul	far

Continued on next page

Table C.11 – continued from previous page

Noun	Semantic variable	Novel Word	Meaning
torch	long	ul	far
leg	long	ul	far
ladder	long	ul	far
belt	long	ul	far
whip	long	ul	far
carrot	long	ul	far
tail	long	ul	far
thread	long	ul	far
mattress	flat	ne	far
newspaper	flat	ne	far
notice	flat	ne	far
painting	flat	ne	far
paper	flat	ne	far
photo	flat	ne	far
postcard	flat	ne	far
poster	flat	ne	far
receipt	flat	ne	far
red card	flat	ne	far
map	flat	ne	far
handkerchief	flat	ne	far
banknote	flat	ne	far
label	flat	ne	far
Christmas card	flat	ne	far
stamp	flat	ne	far
compact disc	flat	ne	far

Table C.12 Testing materials for the 'Language-Specific distributional cues (English)' simulations

Noun	Semantic variable	Correct	Incorrect	Meaning
banana	long	ul	ne	far
celery	long	ul	ne	far

Continued on next page

Table C.12 – continued from previous page

Noun	Semantic variable	Correct	Incorrect	Meaning
sausage	long	ul	ne	far
lamp post	long	ul	ne	far
branch	long	ul	ne	far
cheque	flat	ne	ul	far
blanket	flat	ne	ul	far
bookmark	flat	ne	ul	far
tablecloth	flat	ne	ul	far
envelope	flat	ne	ul	far
cigarette	long	gi	ro	near
pipe	long	gi	ro	near
rope	long	gi	ro	near
ladder	long	gi	ro	near
belt	long	gi	ro	near
map	flat	ro	gi	near
banknote	flat	ro	gi	near
Christmas card	flat	ro	gi	near
stamp	flat	ro	gi	near
compact disc	flat	ro	gi	near

C.5 Stimuli from Saalbach & Imai (2007) (p. 152)

Table C.13 Stimuli used in Saalbach & Imai (2007)

Target	Classifier	Same classifier	Taxonomic	Thematic	Control
Comb	Ba	Key	Hair dryer	Hair	Ticket
Pistol	Ba	Umbrella	Canon	Bullet	Stamp
Scissors	Ba	Fan	Cutter	Paper	TV
Chain	Tiao	Carp	Rope	Lock	Poster
Necklace	Tiao	Blanket	Ring	Dress	Book
Towel	Tiao	Eel	Handkerchief	Shower	Potato
Mountain	Zuo	Tower	Hill	Snow	Necklace
Bell	Zuo	Building	Buzzer	Temple/church	Bike
Piano	Jia	Ladder	Violin	Music book	Scarf
Plane	Jia	Swing	Boat	Airport	Chain
Flower	Duo	Cloud	Tree	Vase	Cup
Newspaper	Zhang	Bed	Book	Morning	Tube
Drum	Mian	Wall	Trumpet	Sticks	Scissors
Tent	Ding	Hat	Sleeping bag	Campfire	Table