Statistical language learning

Luca Onnis

Thesis submitted for the degree of Doctor of Philosophy

Department of Psychology

University of Warwick

October 2003

Table of contents

List of Figures ..... *vi*

List of Tables ..... *viii*

Acknowledgements ..... *ix*

Declaration ..... *x*

Abstract ..... *xi*

Chapter 1 ..... *1*

Statistical language learning **2**

Chapter 2 ..... *14*

Detecting non-adjacent structural dependencies in language **15**

Detection of invariant structure through context variability **17**

Testing the zero-variability hypothesis **20**

Experiment 1 ..... **21**

Method 21
Participants 21
Materials 21
Procedure 22
Results and Discussion 23

General discussion **24**

Chapter 3 ..... *30*

The Variability effect: A graded, associative account **31**

Simulation 1 – The Variability Effect Hypothesis **34**

Method 37
Networks 37
Materials 37
Procedure 37
Results and Analyses 38

Simulation 2 - The zero-variability hypothesis **41**

Method 41
Networks 41
Materials 42
Procedure 42
Results and Analyses 42

Learning nonadjacent structure in SRNs **43**

Conclusions **50**

Chapter 4 ..... *52*

The Variability effect across modalities 53

Experiment 2 - Visual Sequential (VS) version 56

Method 56
Participants 56
Materials 56
Procedure 56
Results and discussion 57

The confirmatory bias in the Variability Experiments 59

Experiment 3 - Visual Sequential Abridged version (VSA) 62

Method 62
Participants 62
Materials 62
Procedure 63
Results and discussion 63

Experiment 4 - Visual Temporal (VT) version 65

Method 65
Participants 65
Materials 65
Procedure 65
Results and discussion 66

General Discussion 68

Chapter 5 ..... *71*

Bootstrapping abstract linguistic representations 72

Generalisation under conditions of variability 75

Experiment 5 (Human data) 83

Method 83
Participants 83
Materials 83
Procedure 84
Results and discussion 84

Simulation 3 (SRN data) 85

Method 86
Networks 86
Materials 86
Procedure 87
Results and Analyses 87

General Discussion 89

Chapter 6 ..... *95*

The debate over the nature of linguistic representations 96

Are algebraic and statistical computations empirically separable? 98

Experiment 6 106

Method 106
Participants 106
Materials and design 106
Procedure 107
Results 108
Discussion 108

Experiment 7 110

Method 110
Participants 110
Materials and design 110
Procedure 111
Results 111
Discussion 111

Experiment 8 113

Method 113
Materials and design 113
Procedure 114
Results 114
Discussion 114

Experiment 9 117

Method 117
Participants 117
Materials and design 117
Procedure 118
Results 118
Discussion 118

Experiment 10 120

Method 120
Participants 120
Materials and design 120
Procedure 121
Results 121
Discussion 121

Experiment 11 123

Method 123
Participants 123
Materials and design 123
Procedure 124
Results and discussion 124

Experiment 12 125

Method 125
Materials and design 125
Procedure 125
Results 126
Discussion 126

Experiment 13 128

Method 128
Participants 128
Materials and design 128

Procedure 129
Results 129
Discussion 129

General discussion 131

Chapter 7 143

Recovery from overgeneralizations in language acquisition 144

Baker's Paradox and linguistic quasi-productivity 146

The logical problem of language acquisition 149

Learning Argument Structure: semantic bootstrapping 150
Learning Argument Structure: Construction Grammar 154
Learning Argument Structure from non-occurrence 155

Causative alternations in child-directed speech 157

Simplicity and Language 161

Modeling language learning with simplicity 162

The Models 163
Calculating Code-Length for each element 164
Simulating recovery from overgeneralization with an artificial language 168

Conclusions and future directions 171

Chapter 8 173

Acquisition and Evolution of quasi-regular languages: Two puzzles for the price of one 174

The logical problem of language evolution 175

Simplicity-Based Language Learning: The Learner as Gambler 175
Learning a rudimentary language 177
Language Learning over Generations - ILM simulations 182
Results 184

Discussion and conclusion 186

Chapter 9 192

Discussion and conclusions 193

Limits and future directions 196

Extensions to the variability effect 196
What is learnt in Artificial Grammars 199
Solving the language acquisition and evolution puzzles with Artificial Grammars 203

References 206

APPENDIX A 222

List of Figures

Figure 1. Total percentage endorsements from Gómez (2002) for the different conditions of variability of the middle item. 19
Figure 2. Total percentage endorsements in Experiment 1 for different variability. 24
Figure 3. A Simple Recurrent Network (after Elman, 1990) 36
Figure 4. Percent accuracy for Simulation 1 across 4 conditions of variability. 40
Figure 5. U-shape learning curve in SRNs for Simulation 2. Error bars are SEM. 42
Figure 6. MDS analysis of hidden unit trajectories. A network trained on 2 Xs fails to achieve the needed separation: all 6 trajectories remain close to each other all the way through the end of training. Hence the network can never form correct predictions of the successor to the X. 48
Figure 7. MDS analysis of hidden unit trajectories in the 24X condition: all 6 trajectories start out, on the left side, from the same small region, and progressively diverge to result in three pairs of two representations. 48
Figure 8. MDS analysis for a network trained on 1 X. Like in the 24X case, the network is successful in separating out the corresponding internal representations: The terminal points of each trajectory end up in different regions of space. 49
Figure 9. Total percentage endorsements in Experiment 2 for different variability. 58
Figure 10. Percent correct responses for Experiment 3 64
Figure 11. Percent correct responses for Experiment 4 67
Figure 12. Percent accuracy in generalising to a new embedding across 3 conditions of variability: null, small, and large. 85
Figure 13. Results from Simulation 3 on generalisation to new embeddings plotted against results obtained experimentally in Experiment 5. 88
Figure 14. At the top of the frame, a sample of the training speech is shown, with "words" shown in different colours and part-words underlined. Underneath, is a sample of a test pair: In Experiments 14-16, words were compared to part-words, in 17-20, rule-words were compared to part-words. The results for each participant, in terms of percentage preference for part-word or word/rule-word, is represented by a dot. The mean for all participants is indicated above a vertical line. Experiment 6 – segmentation task. 109
Figure 15. Experiment 7 – segmentation task with randomized phonology. 112
Figure 16. Experiment 8 – segmentation task with no structure. 116
Figure 17. Experiment 9 – generalization task. 119
Figure 18. Experiment 10 – generalization task with randomized phonology. 122
Figure 19. Experiment 11 – generalization task with gap 124
Figure 20. Experiment 12 – generalization task with gap and randomized phonology. 127
Figure 21. Experiment 13 – generalization task with gap and no structure. 130
Figure 22. Comparison between a traditional ALL task (above) and the segmentation task used by Peña et al. (below). 140
Figure 23. Comparison between the ALL task used by Gómez (2002) with large variability of middle items (above) and a hypothetical mirror segmentation task (below), where low-transitional probabilities between the As and the Xs would lead to wrong segmentation. 141
Figure 24. Comparison between the ALL task used In chapter 2 with no variability of middle items (above) and a hypothetical mirror segmentation task (below), where unwanted nonadjacent dependencies between the Xs and the As having relatively high conditional probabilities would lead to an impossible task 142
Figure 25. The structure of the toy language mimics that of Baker's Paradox for alternations. a_1 and a_2 could be blocked from occurring in BA and AB constructions respectively by entries in the exceptions element such as a_2b_1, a_2b_2 or b_1a_1, b_2a_1 etc. For the first generation agent in each simulation, however, all As occurred in both contexts (that is, they were 'alternating'). 'Cut', 'fall', and 'break' are examples of alternating and non-alternating verbs. Levin (1993) provides an extensive list of alternations in English. 178
Figure 26. The codelength (number of bits) associated with each hypothesis grammar entertained by a learner after exposure to 50 sentences of a language containing 11 exceptions. The shortest codelength is obtained by the 12th hypothesis, i.e. the one containing 11 exceptions. (the first

contains none, being completely regular). Although it is not obvious from the figure, the 12th hypothesis specifies exactly the 11 exceptions contained in the language. *180*
Figure 27. The number of exceptions contained in the language transmitted by each learner to the subsequent generation in four simulations with differing corpus sizes. Where the number of exceptions was stable across several generations, for example seven exceptions in c) or the final 600 generations of d), the sentences specified as exceptions were the same for each generation. It is important to note the difference in scale for number of exceptions for a), b), c) and d). *185*

List of Tables

Table 1. Percentage of endorsements for trained versus untrained strings and total accuracy in each of the five set-size conditions. 58
Table 2. Positive bias for the Visual Sequential experiment 60
Table 3. Positive bias for the Auditory experiment. 60
Table 4. Percent correct responses for Experiment 4 expressed in terms of seen (trained) and unseen (untrained) items recognised correctly 67
Table 5. Percentage of words beginning with each consonant for syllables in initial/medial/final word position in Peña et al.'s studies 103
Table 6. Summary of the design of the experiments. The first column lists the Experiment, the second column lists the experiment number in Peña et al.'s study. "Syllable positions" indicates whether syllables occurred in the original initial/medial/final positions from Peña et al. The "Structure" column indicates whether the language contained nonadjacent dependencies or not, and the effect indicates the statistical result (indicates that there was a significant reverse effect, i.e., there was a preference for part-words over rule-words in Experiment 10).* 105
Table 7. Alternating and non-alternating verbs across contexts. 157
Table 8. Verbs in child-directed speech occurring in transitive and intransitive contexts pooled from the CHILDES English sub-corpora (MacWhinney, 2000). 160
Table 9. Code-lengths of Models 1 and 2 for successively large corpora. Code-lengths in bold show the shorter codes for the corpus size. 170
*Table 10. Sentences allowable under [3]. Rows are first words, columns are second words. The re-write rules license half the sentences in this table; blocked sentences are denoted *. The learner was able to discover exceptions to the rules such as a_2 appearing in first position or a_1 appearing in second position.* 182

Acknowledgements

I would like to thank Nick Chater for his indefatigable support and scholarly advice throughout the last three years. Nick has been a vital and constant source of reference and stimulation, while at the same time allowing me to pursue a personal and unconstrained research path in the directions that most suited my intellectual thirst and curiosity. He has also been unconditionally supportive for many practical aspects of my English life at Warwick. I also feel deeply indebted to Morten Christiansen (Cornell University), who since my second year has fuelled me with invigorating challenges and has taught me many skills.

Most of the work presented here is the fruit of daily collaborations, filtered coffees and Sunday countryside walks with Matthew Roberts and Padraic Monaghan. We have learnt together and together we have become friends. I will miss them both enormously.

Several other scholars deserve mention for sharing their thoughts and skills: Axel Cleeremans (ULB, Brussels), Rebecca Gómez (Arizona), Arnaud Destrebecq (ULB, Brussels), and Bob French (Liege).

This work is dedicated to the memory of my grandfather Adamo Volpe (1917-2003), who lived his life fully until the very last minute.

This thesis was supported by European Union Project HPRN-CT-1999-00065, "Basic mechanisms of learning and forgetting in natural and artificial systems".

Declaration

I hereby declare that the research reported in this thesis is my own work unless otherwise stated. No part of this thesis has been submitted for a degree at another university.

Parts of chapter 2 and 4 have been published in Onnis, Christiansen, Chater, & Gómez, (2003). The contents of chapter 3 form part of Onnis, Destrebecq, Christiansen, Chater, & Cleeremans (submitted). Chapter 6 as been submitted for publication in Onnis, Monaghan, Chater, & Richmond. Material in chapter 7 is published as Onnis, Roberts, & Chater (2002), while parts of chapter 8 are to be published in Roberts, Onnis, & Chater (in press).

The MDS analyses and graphs in chapter 3 were provided by Axel Cleeremans. The computer scripts for the simplicity simulations and the equations reported in chapters 7, 8, and Appendix A were written by Matthew Roberts.

Luca Onnis

Abstract

Theoretical arguments based on the "poverty of the stimulus" have denied a priori the possibility that abstract linguistic representations can be learned inductively from exposure to the environment, given that the linguistic input available to the child is both underdetermined and degenerate. I reassess such learnability arguments by exploring a) the type and amount of statistical information implicitly available in the input in the form of distributional and phonological cues; b) psychologically plausible inductive mechanisms for constraining the search space; c) the nature of linguistic representations, algebraic or statistical. To do so I use three methodologies: experimental procedures, linguistic analyses based on large corpora of naturally occurring speech and text, and computational models implemented in computer simulations.

In Chapters 1, 2, and 5, I argue that long-distance structural dependencies – traditionally hard to explain with simple distributional analyses based on n-gram statistics - can indeed be learned associatively provided the amount of intervening material is highly variable or invariant (the Variability effect). In Chapter 3, I show that simple associative mechanisms instantiated in Simple Recurrent Networks can replicate the experimental findings under the same conditions of variability. Chapter 4 presents successes and limits of such results across perceptual modalities (visual vs. auditory) and perceptual presentation (temporal vs. sequential), as well as the impact of long and short training procedures. In Chapter 5, I show that generalisation to abstract categories from stimuli framed in non-adjacent dependencies is also modulated by the Variability effect. In Chapter 6, I show that the putative separation of algebraic and statistical styles of computation based on successful speech segmentation versus unsuccessful generalisation experiments (as published in a recent *Science* paper) is premature and is the effect of a preference for phonological properties of the input. In chapter 7 computer simulations of learning irregular constructions suggest that it is possible to learn from positive evidence alone, despite Gold's celebrated arguments on the unlearnability of natural languages. Evolutionary simulations in Chapter 8 show that irregularities in natural languages can emerge from full regularity and remain stable across generations of simulated agents. In Chapter 9 I conclude that the brain may endowed with a powerful statistical device for detecting structure, generalising, segmenting speech, and recovering from overgeneralisations. The experimental and computational evidence gathered here suggests that statistical language learning is more powerful than heretofore acknowledged by the current literature.

Chapter 1

Statistical language learning

To what extent is language learnable from experience? Does the information available to the child in the form of statistical regularities allow learning core aspects of language such as syntactic structures, segmenting speech, generalising and recovering from overregularisations? The remarkable speed and apparent implicitness with which human infants acquire a language in their first years of life has lead many theorists to dismiss a priori the idea that statistical information inherent in the language plays a central role in acquisition. Theoretical arguments based on the 'poverty of the stimulus' (Gold, 1967; Chomsky, 1965; Pinker, 1984) have drawn attention to the fact that positive evidence available to the learner is insufficient to distinguish between grammatical and ungrammatical utterances, and that online speech is full with flaws and missing elements. Because the target language seems both underdetermined and degenerate, successful learning must occur *despite* the nature of the input on a deductive basis by means of an innate mental language system.

This thesis takes on a recent and renewed interest in the analysis of language acquisition from an inductive perspective, and tries to assess empirically and computationally what can be learned from the environment. We can broadly term this field statistical language learning. Core issues tackled in this area are: (a) how reliable is statistical information for bootstrapping linguistic structure in the form of low-level prosodic, phonological, and distributional cues? (b) In the face of a combinatorial explosion of potentially valid hypotheses about some linguistic structure given the cues in the input, what psychologically plausible constraints should apply to the learning device? For

instance, Redington, Chater, & Finch (1998) pointed out that a totally unconstrained search with n items and m syntactic categories (where each item belongs to a single syntactic category and assuming the number of categories is known a priori), would imply considering m^n possible mappings, and that there are already more than a million permutations with only 20 items and 2 syntactic categories. From an empiricist point of view this task is even harder because the number of syntactic categories is not innately specified. Clearly, statistical analyses that entertain all possible relations among words would be computationally intractable; (c) Does language learning ultimately necessitates a language-specific device or Universal Grammar, or does it impinge on general-purpose mechanisms that support human learning broadly? As a result of a shift to nativism in American linguistics towards the late 1950s, the role of inductive learning - what can be learned from the environment given general-purpose inductive mechanisms - has been downplayed as not powerful enough. Recently, various researchers have started to reassess empirically and computationally both the amount of information inherently available in the linguistic input and the power and types of mechanisms that might be plausibly engaged in language learning; (d) What is the nature of linguistic representations in the brain – algebraic-like or statistical?

The field of language acquisition has recently benefited from a wave in computational modeling. Computational models can be seen as intermediate tools that mediate between a purely "verbal" theory and a purely experimental paradigm (Broeder & Murre, 2003). As a computer implementation of a theory a computational model requires the modeller to make more explicit the assumptions underpinnings her theory. Because it involves an input, a process,

and an output, it can also be subjected to experimental manipulations that test different conditions of behaviour. As an intermediate between theory and experiment, a model can thus be judged in terms of how well it implements the theory as well as how well it fits the data gathered. In this thesis computational models are coupled with experimental paradigms in order to accumulate more robust evidence about a given issue. In this work I specifically focus on four related aspects of language learning from experience: detecting nonadjacent invariant structure, generalising beyond experience to novel instances given an invariant structure, segmenting speech into core constituents, and recovering from overgeneralisations. Detecting invariant structure and generalising are seen by many as the hallmark of discovering syntactic structure in language (Chomsky, 1957). Research on statistical learning in adults and infants has shown that humans are particularly sensitive to statistical properties of the input, for instance, transitional n-gram probabilities. Although this may help children segment speech (Saffran, Aslin, & Newport, 1996) it has been argued, however, that this source of information may not help in detecting nonadjacent dependencies, in the presence of substantial variability of the intervening material (Gómez, 2002). Words in the language are organised into constituents called phrases, groupings of words that behave as units (typical constituents are Noun Phrases, Verb Phrases, Prepositional Phrases, Adjective Phrases). The position of such constituents is not fixed in a sentence because of the recursivity of syntax: for instance, a Noun Phrase constituent that contains a Prepositional Phrase can in turn contain another Noun Phrase. Recursivity generates non-local dependencies, the fact that two words can be syntactically dependent even

though they occur far apart in a sentence. Consider subject-verb agreement in English in the following examples:

(1) Mark runs, She runs, The rabbit runs

(2) John and Mark run, The rabbit with the white fur runs

(3) The woman with the blue dress is kind

(4) The women with the blue dress are kind

As one can see, a near-neighbour analysis such as **Mark run* in (2) or *the blue dress is kind* in (4) does not yield the correct structural dependency. In Chapter 2, in particular, I discuss that detecting long-distance relationships like verb-noun agreement and tensed verbs are hard to explain in terms of simple distributional analyses based on n-gram statistics such as transitional probabilities. This is because the intervening material is extremely variable and hence has to be ignored for the structural constraints to be learned. Sequences in natural languages typically involve some items belonging to a relatively small set (functor words and morphemes like *am, the, -ing, -s, are*) interspersed with items belonging to a very large set (e.g. nouns, verbs, adjectives). Crucially, this asymmetry translates into patterns of highly invariant nonadjacent items separated by highly variable material (*am cooking*, *am working*, *am going*, etc.). On the other end, nonadjacent contingencies such as number agreement may share the very same embedded material: consider sentence (1) versus (2) below:

(5) The book on the shelf is dusty

(6) The books on the shelf are dusty.

In either case - large variability or no-variability of intervening items - knowledge of n-gram conditionals cannot be invoked for detecting invariant structure. The same chapter hence introduces the Variability Effect hypothesis, in which I empirically show that learners are better at detecting long-distance dependencies with either zero or high variability. I show a U-shape in learning long-distance contingencies as a function of the number of intervening items. Gómez (2002) has proposed that alternative sources of information may be attended to simultaneously by learners. With several potential cues in competition, human learning seems extremely flexible and naturally biased toward the most informative ones in an attempt to maximally reduce uncertainty.

In chapter 3, I discuss the extent to which simple associative mechanisms instantiated in connectionist models can account for the Variability Effect. A Simple Recurrent Network (SRN) is able to detect nonadjacent sequential contingencies by developing graded representations in hidden units that simultaneously maintain similarities and differences between several sequences. Crucially this happens in the presence of either zero variability or large variability, thus replicating the U-shape pattern obtained experimentally.

Chapter 4 examines the extent to which a U-shape learning curve attributed to the Variability Effect is modality-independent and may be affected by training length. In two new experiments the same training and test stimuli used in chapter 2 were presented visually on a computer screen. The obtained U-shape curve is less marked when whole sentences appear on the screen. One possible explanation is that attending to visually presented stimuli is less demanding cognitively or makes the structure stand out visually, explaining the

ceiling effect. In another experiment, presenting words one by one on the screen (thus mirroring the sequential presentation of the auditory version) yields results that are at the same time surprising and difficult to interpret, as the U-shape turns into an S-shape. In a third experiment, new participants were administered the same auditory experiment of chapter 2 with a halved training regime. This manipulation was initially motivated by the desire to reduce the large variation in scores between subjects within each condition, on the assumption that 20 minutes of training might produce boredom or distraction in participants. However, the U-shape did not emerge with 10 minutes of training exposure. Overall, the chapter tackles the limits of interpretability of single ALL results and cautions against drawing fast conclusions without a good battery of tests. In the AGL community it is often believed that because of their artificiality and abstractness artificial grammars capture the essence of learning at a highest, indeed abstract way. The results presented here point to different performance results depending on the training regime and the way the stimuli are perceptually perceived. The issue is explored further in Chapter 6 when phonological confounds are shown to explain away strong theoretical claims about the separability of statistical and algebraic computations.

Generalisation is regarded as a core aspect of linguistic knowledge (Chomsky, 1957): although learners are exposed to a limited amount of language they produce an infinite number of sentences in their life. The ability to abstract beyond exemplars encountered is thus a critical feature of syntax acquisition. Chapter 5 discusses generalisation in the light of the variability results. Whereas the experiments in Chapter 3 test preference for grammatical items previously encountered in the training phase, in Chapter 5 I test empirically whether

generalisation to *novel* stimuli is supported under the same conditions of variability involved in detecting invariant structure.

Chapter 6 deals with speech segmentation and generalisation. The speech signal is mostly continuous and word boundaries are rarely marked by acoustic cues such as pauses. This poses a serious inferential problem to the child who lacks knowledge of the syntax and semantics of the language as well as the phonological properties of the lexicon. Here I discuss segmentation strategies with specific reference to an article by Peña, Bonatti, Nespor, and Mehler (2002). Many theories of language acquisition debate whether processing is dependent on statistical computations alone or whether it needs algebraic computations. Peña *et al.* recently argued that speech segmentation was a statistical process, whereas learning structural generalizations was due to algebraic computations. In a series of experiments, extending those by Peña *et al.*, I found that participants had strong preferences for phonemes in certain utterance positions. I found no evidence for the statistical/algebraic distinction: the results from Peña *et al.* were a consequence of the impact of phonological preferences on language processing. I reassess the debate on algebraic versus statistical computation in the light of the obtained results. Chapter 6 ties in well with the previous ones for two reasons: firstly, they deal with the issue of exploiting long-distance dependencies for segmenting speech and generalising to novel items, thus adding another piece to the puzzle. Secondly, they elaborate on the methodological considerations started in chapter 5 about the perceptual non-neutrality of artificial stimuli, which is often incorrectly taken for granted.

Chapters 7 and 8 conclude the statistical explorations into language by looking at the other side of generalisation, namely how a learner can recover

from overgeneralisations which are known to be spontaneously generated by children (such as **I disappeared the rabbit*) without direct negative evidence, i.e. without direct correction from the caretaker. This is a general problem of inductive inference. Overgeneralizations are a common feature of language development. In learning the English past tense, children typically overgeneralize the '-ed' suffix marker, producing constructions such as **we holded the baby rabbits* (Pinker, 1995). Language learners recover from these errors, in spite of the lack of negative evidence and the infinity of allowable constructions that remain unheard (Gold, 1967). It has been argued that this favours the existence of a specific language-learning device (e.g. Chomsky, 1980; Pinker, 1989). This is an aspect of the 'Poverty of the Stimulus' argument. I report on a statistical model of language acquisition, which suggests that recovery from overgeneralizations may proceed from positive evidence alone. Specifically, I show that adult linguistic competence in quasi-regular structures may stem from an interaction between a general cognitive principle, *simplicity* (Chater, 1996) – based on the mathematical theory of Kolmogorov Complexity (Kolmogorov, 1965) – and statistical properties of the input. Under what is known as Baker's Paradox (Baker, 1979) non-occurrence in the input is not in itself evidence for the incorrectness of a construction because an infinite number of unheard sentences are still correct. One type of irregularities that Baker referred to can be broadly labeled *alternations* (Levin, 1993; see also Culicover, 2000). For instance, the dative alternation in English allows a class of verbs to take both the double-object construction (*He gave Mark the book*) and the prepositional construction (*He gave the book to Mark*). Hence the verb *give* alternates between two constructions. However, certain verbs seem to be constrained to one possible

construction only (*He donated the book to Mark* is allowed, whereas **He donated Mark the book* is not). Such verbs are non-alternating. From empirical studies we know that children do make overgeneralization errors that involve alternations, such as **I said her no* (by analogy to *I told her no*, Bowerman, 1996; Lord 1979).

In chapter 7, I present alternation phenomena from the CHILDES database (MacWhinney, 2000) of child-directed speech which will be used in the computer model. The simplicity principle (Chater, 1996; Chater & Vitányi, 2001) states that the cognitive system seeks the hypothesis that provides the briefest representation of the available data – here the linguistic input to the child. This model allows learning from positive evidence alone in a probabilistic sense, contrasting with Gold's (1967) negative theorems. Data gathered from the CHILDES database were used as an approximation of positive input the child receives from adults. I consider linguistic structures that would yield overgeneralization. Two computer simulations incorporating simplicity were run corresponding to two different hypotheses about the grammar: (1) The child assumes that there are no exceptions to the grammar. This hypothesis leads to overgeneralization. (2) The child assumes that some constructions are not allowed. By measuring the cost to encode a structure given its probability P of occurrence as $\log_2(1/P)$, the second hypothesis was preferred as it lead to a shorter input description and eliminated overgeneralization.

While chapter 7 attempts to solve the long-debated logical problem of language acquisition, chapter 8 takes an evolutionary perspective. The relative diachronic stability of quasi-productive constructions in linguistic codes poses a puzzle for accounts based on the principle of parsimony of representation. The

logical problem of language evolution is concerned with how quasi-regularities such as alternations could have possibly emerged in natural languages and why they were not eliminated over generations, if these constituted a serious learning problem. In particular, I consider the fact that languages are never fully productive, although full productivity would be a desirable solution in terms of learnability over generations (Kirby, 2001; Hurford 2000). I present several simulations charting the emergence and stability of irregularities across 1000 generations of artificial simplicity-based learners using an artificial language. In all simulations grammar induction is by simplicity. Randomly set proto-grammars are transmitted across 1,000 generations of communicating agents. At each generation a simplicity learner seeks the shortest representation of the available data. As a result, overgeneral grammars are not handed down over the next generation and alternations remain stable, at least across a number of generations.

Finally, Chapter 9 pulls the lines of research on statistical language learning together, discussing the merits and limits of a distributional approach. I hope to show that beyond well-founded theoretical claims for the unlearnability of language in some deep sense, there is ample scope for setting a rigorous research agenda for evaluating experimentally and computationally what aspects of language can be learned from experience and what cannot. The relative recency of the area of statistical language learning as well as the preliminary nature of the investigations reported here can only warrant a cautionary position that eschews polarized views. Ultimately, it is suggested here that the human brain may be endowed with a powerful statistical device for detecting structure,

generalising, segmenting speech and recovering from overgeneralisations found in natural languages.

This work is exploratory by necessity because none of the studies that I report can claim a definitive answer to a specific issue, although they are all self-contained projects that have been published or submitted for publication. I also perform a cursory exploration in language learning in as much as the experiments and simulations reported here do not deal with real linguistic utterances in real communicative contexts, but rather make use of simplified grammars technically known as artificial or finite-state grammars. The virtues of such a simplification will soon result apparent, particularly for the possibility of carefully controlling the conditions of learning in experimental settings, as well as making learning a computationally tractable issue in computer simulations. Using artificial language stimuli enables precise control over the learning environment, and allows systematic manipulation and testing of specific structures. As we shall see, artificial stimuli need not be entirely abstracted from real languages: both the experimental stimuli and the computer simulations reported here are empirically motivated by statistical analyses of large corpora of real language, such as the CHILDES database and the British National Corpus.

The reader may also be struck to note that, although I deal with language acquisition throughout this work, none of the experiments involve infants or children. This is certainly a caveat. In recent times, insights and methodologies from two lines of research have been combined: one involving studies of artificial grammar learning (henceforth AGL) in adults (e.g. Reber, 1967, 1969; Morgan & Newport, 1981; Valian & Levitt, 1996) and another examining infant learning examining infant learning of artificial language material (ALL). Because

the two areas are now beginning to be merged, and because the learning that results from adults is better understood, it is customary to gather preliminary data from adult performance as a baseline against infant performance to be tested later.

Chapter 2

Detecting non-adjacent structural dependencies in language

Research in artificial grammar learning (AGL) and artificial language learning (ALL) in infants and adults has revealed that humans are extremely sensitive to the statistical properties of the environment they are exposed to. This has opened up a new trend of investigations aimed at determining empirically the processes involved in so-called statistical learning.

Several mechanisms have been proposed as the default that learners use to detect structure, although crucially there is no consensus in the literature over which is most plausible or whether there is a default at all. Some researchers have shown that learners are particularly sensitive to transitional probabilities of bigrams (Saffran, Aslin, & Newport, 1996): confronted with a stream of unfamiliar concatenated speech-like sound they tend to infer word boundaries between two syllables that rarely occur adjacently in the sequence. Sensitivity to transitional probabilities seems to be present across modalities, for instance in the segmentation of streams of tones (Saffran, Johnson, Aslin, and Newport, 1999) and in the temporal presentation of visual shapes (Fiser & Aslin, 2002).

Other researchers have proposed exemplar- or fragment-based models, based on knowledge of memorised chunks of bigrams and trigrams (Dulany *et al.*, 1984; Perruchet & Pacteau, 1990; Servan-Schreiber & Anderson, 1990) and learning of whole items (Vokey & Brooks, 1992). Yet others have postulated rule-learning in transfer tasks (Reber, 1967; Marcus, Vijayan, Rao & Voshton, 1999). In addition, knowledge of chained events such as sentences in natural languages

require learners to track nonadjacent dependencies where transitional probabilities are of little help (Gómez, 2002).

In this Chapter I propose that there may be no default process in human sequential learning. Instead, learners may be actively engaged in search for good sources of reduction in uncertainty. In their quest, they seek alternative sources of predictability by capitalizing on information that is likely to be the most statistically reliable. This hypothesis was initiated by (Gómez, 2002) and is consistent with several theoretical formulations such as reduction of uncertainty (Gibson, 1991) and the simplicity principle (Chater, 1996), which states that the cognitive system attempts to seek the simplest hypothesis about the data available. Given performance constraints, the cognitive system may be biased to focus on data that will be likely to reduce uncertainty as far as possible[1]. Specifically, whether the system focuses on transitional probabilities or non-adjacent dependencies may depend on the statistical properties of the environment that is being sampled. Therefore, by manipulating the statistical structure of that environment, it is perhaps possible to investigate whether active search is at work in detecting structure.

In two experiments, I investigated participants' degree of success at detecting invariant structure in an AGL task in 5 conditions where the test items and test task are the same but the probabilistic environment is manipulated so as to change the statistical landscape substantially. I propose that a small number of alternative statistical cues might be available to learners. I aim to show that, counter to intuition, orthogonal sources of reliability might be at work in different

experimental conditions leading to successful or unsuccessful learning. I also asked whether my results are robust across perceptual modalities by running two variations of the same experiment, one in the auditory modality and one in the visual modality. My experiments are an extension of a study by Gómez (2002), which I first introduce.

Detection of invariant structure through context variability

Many sequential patterns in the world involve tracking nonadjacent dependencies. For example, in English auxiliaries and inflectional morphemes (e.g., *am cooking, has travelled*) as well as dependencies in number agreement (*the books on the shelf are dusty*) are separated by various intervening linguistic material. One potential source of learning in this case might be embedding of first-order conditionals such as bigrams into higher-order conditionals such as trigrams. That learners attend to n-gram statistics in a chunking fashion is evident in a number of studies (Schvaneveldt & Gómez, 1998; Cohen, Ivry, & Keele, 1990). In the example above chunking involves noting that *am* and *cook* as well as *cook* and *ing* are highly frequent and subsequently noting that *am cooking* is highly frequent too as a trigram. Hence we may safely argue that higher order n-gram statistics represent a useful source of information for detecting nonadjacent dependencies. However, sequences in natural languages typically involve some items belonging to a relatively small set (functor words and morphemes like *am, the, -ing, -s, are*) interspersed with items belonging to a very large set (e.g. nouns, verbs, adjectives). Crucially, this asymmetry translates into patterns of highly invariant

[1] We assume that this process of selection is not necessarily conscious, and might for example involve

nonadjacent items separated by highly variable material (*am cooking*, *am working*, *am going*, etc.). Gómez (2002) suggested that knowledge of n-gram conditionals cannot be invoked for detecting invariant structure in highly variable contexts because first-order transitional probabilities, P(Y|X), decrease as the set size of Y increases. Similarly, second-order transitional probabilities, P(Z|XY), also decrease as a function of set size of X. Hence, statistical estimates for these transitional probabilities tend to be unreliable. Gómez exposed infants and adult participants to sentences of an artificial language of the form *A-X-B*. The language contained three families of nonadjacent pairs, notably A_1—B_1, A_2—B_2, and A_3—B_3. She manipulated the set size of the middle element X in four conditions by systematically increasing the number from 2 to 6 to 12 and 24 word-like elements. In this way, conditional bigram and trigram probabilities decreased as a function of number of middle words. In the test phase, participants were required to subtly discriminate correct nonadjacent dependencies, (e.g. A_2-X_1-B_2) from incorrect ones (*A_2-X_1-B_1). Notice that the incorrect sentences were new as trigrams, although both single words and bigrams had appeared in the training phase in the same positions. Hence the test requires very fine distinctions to be made. Gómez hypothesized that if learners were focusing on n-gram dependencies they should learn nonadjacent dependencies better when exposed to small sets of middle items because transitional probabilities between adjacent elements are higher for smaller than for larger set sizes. Conversely, if learners spotted the invariant structure better in the larger set size, Gómez hypothesized that increasing variability in the context must have led them to disregard the highly variable middle elements. Her

---

distribution of processing activity in a neural network.

results support the latter hypothesis: learners performed poorly with low variability whereas they were particularly good when the set size of the middle item was largest (24 middle items; see Figure 1).

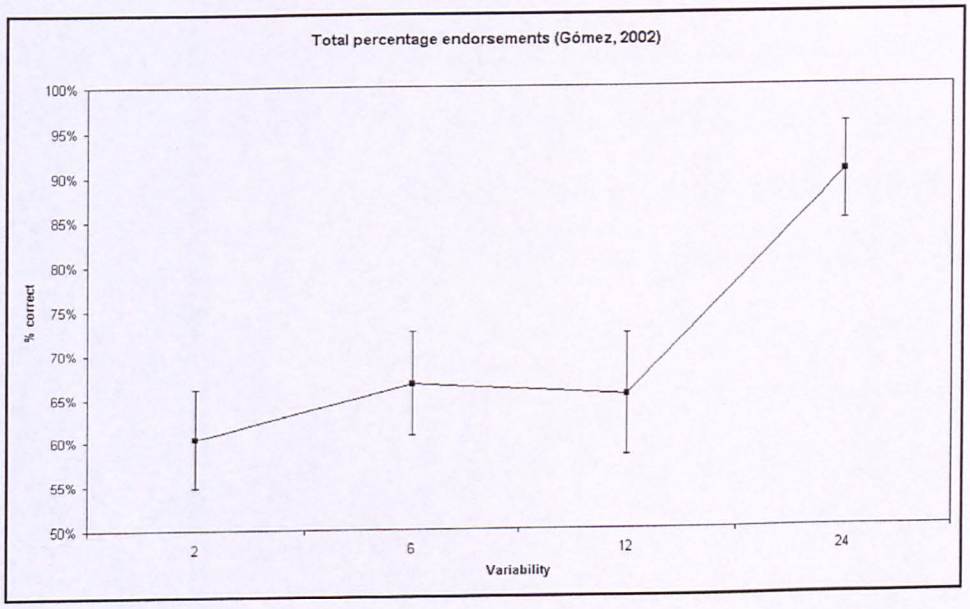

Figure 1. Total percentage endorsements from Gómez (2002) for the different conditions of variability of the middle item.

Testing the zero-variability hypothesis

Gómez proposed that both infant and adult learners are sensitive to change versus non-change, and use their sensitivity to capitalize on stable structure. Learners might opportunistically entertain different strategies in detecting invariant structure, driven by a reduction of uncertainty principle. In this study I am interested in taking this proposal further by exploring what happens when variability between the end-item pairs and the middle items is reversed in the input. Gómez attributed poor results in the middle set sizes to low variability: the *variability effect* seems to be attended to reliably only in the presence of a critical mass of middle items. However, an alternative explanation is that in small set size conditions both nonadjacent dependencies and middle items vary, but none of them considerably more than the other. This may confuse learners, in that it is not clear which structure is non-variant. With larger set sizes middle items are considerably more variable than first-last item pairings, making the nonadjacent pairs stand out as invariant. I asked what happens when variability in middle position is eliminated, thus making the nonadjacent items variable. I replicated Gómez' experiment with adults and added a new condition, namely the zero-variability condition, in which there is only one middle element (e.g. A_3-X_1-B_3, A_1-X_1-B_1). My prediction is that non-variability of the middle item will make the end-items stand out, and hence detecting the appropriate nonadjacent relationships will become easier, increasing mean performance rates.

Intuitively, sampling transitional probabilities with large context variability results in low information gain as the data are too few to be reliable; by the same vein, the lack of variability should produce low information gain for transitional probabilities as well, because it is just obvious what the bigram

structure is. Hence this should make nonadjacent dependencies stand out, as potentially more informative sources of information, by contrast.

The final predicted picture is a U-shape learning curve in detecting nonadjacent dependencies, on the assumption that learning is a flexible and adaptive process.

Experiment 1

Method

Participants

Sixty undergraduate and postgraduate students at the University of Warwick participated and were paid £3 each.

Materials

In the training phase participants listened to auditory strings generated by one of two artificial languages (L1 or L2). Strings in L1 had the form *aXd*, *bXe*, and *cXf*. L2 strings had the form *aXe*, *bXf*, *cXd*. Variability was manipulated in 5 conditions, by drawing X from a pool of either 1, 2, 6, 12, or 24 elements. The strings, recorded from a female voice, were the same that Gómez used in her study and were originally chosen as tokens among several recorded sample strings in order to eliminate talker-induced differences in individual strings.

The elements *a*, *b*, and *c* were instantiated as *pel*, *vot*, and *dak*; *d*, *e*, and *f*, were instantiated as *rud*, *jic*, *tood*. The 24 middle items were *wadim*, *kicey*, *puser*, *fengle*, *coomo*, *loga*, *gople*, *taspu*, *hiftam*, *deecha*, *vamey*, *skiger*, *benez*, *gensim*, *feenam*, *laeljeen*, *chla*, *roosa*, *plizet*, *balip*, *malsig*, *suleb*, *nilbo*, and *wiffle*. Following the design by Gómez (2002) the group of 12 middle elements were drawn from the first 12 words in the list, the set of 6 were drawn from the

first 6, the set of 2 from the first 2 and the set of 1 from the first word. Three strings in each language were common to all five groups and they were used as test stimuli. The three L2 items served as foils for the L1 condition and vice versa. In Gómez (2002) there were six sentences generated by each language, because the smallest set size had 2 middle items. To keep the number of test items equal to Gómez I presented the 6 test stimuli twice in two blocks, randomizing within blocks for each participant. Words were separated by 250-ms pauses and strings by 750-ms pauses.

Procedure

Six participants were recruited in each of the five set size conditions (1, 2, 6, 12, 24) and for each of the two language conditions (L1, L2) resulting in 12 participants per set size. Learners were asked to listen and pay close attention to sentences of an invented language and they were told that there would be a series of simple questions relating to the sentences after the listening phase. During training, participants in all 5 conditions listened to the same overall number of strings, a total of 432 token strings. This way, frequency of exposure to the nonadjacent dependencies was held constant across conditions. For instance participants in set-size 24 heard six iterations of each of 72 type strings (3 dependencies x 24 middle items), participants in set-size 12 encountered each string twice as often as those exposed to set size 24 and so forth. Hence whereas nonadjacent dependencies where held constant, transitional probabilities decreased as set size increased.

Training lasted about 18 minutes. Before the test, participants were told that the sentences they had heard were generated according to a set of rules

involving word order, and they would now hear 12 strings, 6 of which would violate the rules. They were asked to press "Y" on a keyboard if they thought a sentence followed the rules and to press "N" otherwise.

Results and Discussion

An analysis of variance with Set Size (1 vs. 2 vs. 6 vs. 12 vs. 24) and Language (L1 vs. L2) as between-subjects and Grammaticality (Trained vs. Untrained strings) as a within-subjects variable resulted in a main effect of Grammaticality, $F(1,50)=24.70$, $p<.001$, a main Set Size effect, $F(4,50)=3.85$, $p<.008$, and a Language x Set Size interaction, $F(4,50)=2.59$, $p<.047$. I was particularly interested in determining whether performance across the different set-size conditions would result in a U-shaped function. Consistent with my prediction, a polynomial trend analysis yielded a significant quadratic effect, $F(1,50)=5.85$, $p<.05$. In contrast to Gómez (2002), there was not a significant increase between set size 12 and set size 24, $t(22)=.57$, $p=.568$. This leveling off is responsible for a significant cubic effect, $F(1,50)=9.49$, $p<.005$. Figure 2 summarizes total percentage endorsements for correct answers.

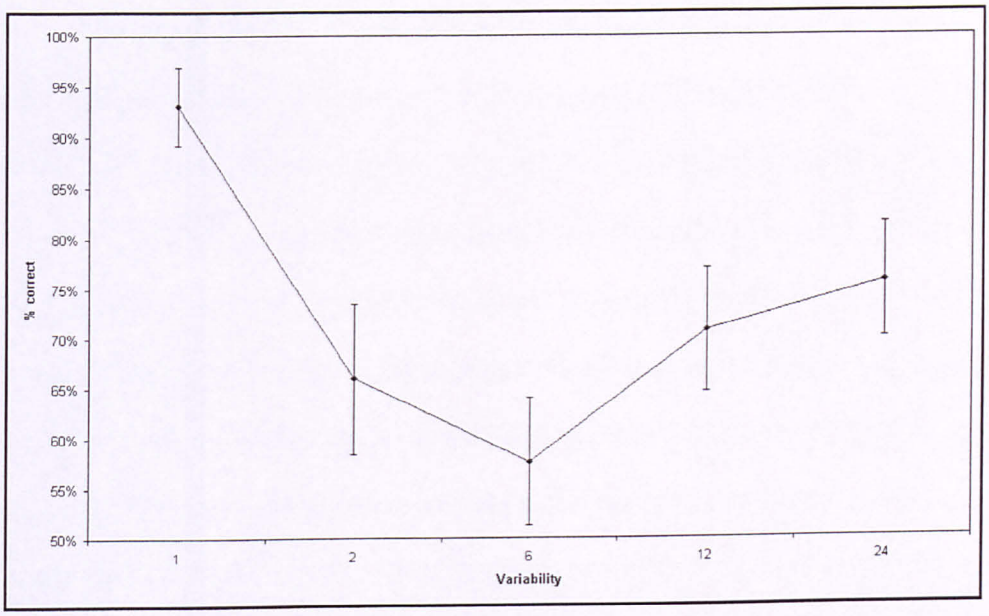

Figure 2. Total percentage endorsements in Experiment 1 for different variability.

General discussion

We used a simple artificial language to enquire into the way learners track remote dependencies. Knowledge of sequence events in the world, including language, involves detecting fixed nonadjacent dependencies interspersed with highly variable material. Gómez (2002) found what I dub a *variability effect*, i.e. a facilitatory effect in detecting invariant structure when the context is highly variable, but not when it is moderately or even little variable. In general, this points to a specific sensitivity to change versus non-change. Conditions 2 to 4 in my Experiment 1 replicate her findings, although performance in terms of percent accuracy seems to improve only gradually from set size 2 to 24, whereas Gómez found a significant difference between set size 12 and 24.

Overall, Gómez' original results do not square well with recent findings of learners' striking sensitivity to n-gram transitional probabilities. Because transitional probabilities are higher in set sizes 2, 6, and 12, performance should be better. Instead, the opposite is the case. I reasoned that perhaps variability in both the middle item and end-point items leave learners in doubt as to what is the invariant structure. Hence, by eliminating variability in the middle item in a new condition, the variability of the nonadjacent items stands out again, this time reversed. However, the effect is, quite counter intuitively, not reversed. Indeed similar performance results are obtained for set size 1 and set size 24. In set size 1 performance is near 100% and significantly better than set size 2 (Experiment 1). One could argue that word trigrams, if recorded perfectly, could suffice to account for performance in set size 1, thus trivializing my results and explaining away the variability effect in this condition. However, as a counter to that it would be reasonable to expect good performance in set size 2 condition too, given the high number of repetitions (72) for only six type strings. A control condition could have been run involving learning six frames (instead of three) with 1 different middle item each (e.g. A_3-X_3-B_3, A_6-X_6-B_6) so as to reproduce the same number of type and token frequencies of set size 2, but with no middle item being *shared* by different frames. However, the doubt of rote learning will be solved in chapter 5, when generalisation to novel middle items will be tested in set size 1.

Similarly, one could argue that good performance in set size 24 could be achieved by strikingly but not impossibly memorizing 72 type strings. However, this would imply good performance in all smaller set sizes as well and this runs counter to data.

Notice also that in all conditions, including set size 1, bigram transitional probabilities by themselves are not sufficient for detecting the correct string *pel wadim rud* from the incorrect one **pel wadim jic* (example taken from L1) as both *pel wadim, wadim rud,* and *wadim jic* appear as bigrams during training. Moreover, Gómez (2002) conjectured that perhaps low discrimination rates in small set sizes are due to overexposure of string tokens during training, resulting in boredom and distraction. My findings disconfirm this hypothesis: if it held true performance would drop even lower in the zero-variability condition, as the type/token ratio decreases even more. Crucially, the finding that there is a statistically significant difference in learning in the two conditions becomes intriguing for several reasons.

The implications of my findings might inform in various degrees both the AGL community and researchers of language development. AGL researchers working mainly with adults have long debated whether there are one or more mechanisms at work in learning structured events from experience. My results suggest that associative learning based on adjacent material may not be the only source of information. There seems to be a striking tendency to detect variant versus invariant structure, and the way learners do it is extremely adaptive to the informational demands of their input. Without claiming exhaustiveness I explored two putative sources of information, namely n-gram transitional probabilities and the variability effect. At this stage I can only give an informal explanation of the reduction of uncertainty hypothesis. Intuitively, sampling bigrams involving middle items under no variability yields no information gain, as the middle item is always the same. Under this condition learners may be driven to shift attention towards nonadjacent structure. Likewise, sampling

bigrams with large variability yields no reduction of uncertainty, as bigram transitional probabilities are very low. In a similar way then, learners may be lead to focus on nonadjacent dependencies. With low variability, sampling bigrams may be reliable enough, hence "distracting" learners away from nonadjacent structure. Other sources may be at work and disentangling the contribution of each of them to learning is an empirical project yet to be investigated. For instance, post-hoc verbal reports from the majority of my participants suggest that, regardless of their performance, they were aware of the positional dependencies of single words in the strings. This piece of information may be misleading for the task: on the one side it reduces uncertainty by eliminating irrelevant hypotheses about words in multiple positions (each word is either initial, middle, or final), on the other side distinguishing *pel wadim rud* from **pel wadim jic* requires more than positional knowledge. I believe that positional knowledge deserves more research in the current AGL literature. Studies of sequential learning have found that it is an important source of information. However, many nonadjacent dependencies are free ranging and hence non-positionally dependent. Further experiments are needed to investigate whether people can detect such non-positionally dependent constraints as *A_x_y_B*, *A_x_y_w_B*, *A_x_y_w_z_B*, equally well.

In the next chapter I will show that these results can be modelled successfully using simple recurrent neural connectionist networks (SRNs) trained in experimental conditions akin to the adult data reported here, obtaining a very similar U-shape curve. SRNs can be thought of as reducing uncertainty in that predictions tend to converge towards the optimal conditional probabilities of observing a particular successive item to the sequence presented up to that point.

The SRNs specific task was to predict the third nonadjacent element B_i correctly. Minimizing the sum squared error maximizes the probability of the next element, given previously occurring adjacent elements (McClelland, 1998). This is equivalent to exploiting bigram probabilities. As we have seen, conditional probability matching only yields suboptimal behaviour. To overcome this, SRNs possess a stack of memory units that help them maintain information about previously encountered material. Crucially, they maintain a trace of the correct non-adjacent item A_i under either no variability or large variability only. This happens by forming separate graded representations in the hidden units for each nonadjacent dependency.

The reduction of uncertainty hypothesis may also be given a formal account in terms of active data selection (MacKay, 1992, Oaksford & Chater, 1994), a form of rational analysis (Anderson, 1990). However, the details of such model are outside the scope of this chapter (see Monaghan, Chater & Onnis, in preparation). Overall, framing my results within a reduction of uncertainty principle should prompt new research aimed at discovering in which carefully controlled statistical environments multiple sources are attended to and either discarded or integrated.

Finally, my findings might inform research in language development. Gómez (2002) found that infants attend to the variability effect. I am currently investigating whether the U-shape curve found in my experiments applies to infant learning as well. The fact that performance in the zero-variability condition is very good is consistent with various findings that children develop productive linguistic knowledge only gradually building from fixed item-based constructions. According to the Verb Island hypothesis for example (for a

review, see Tomasello, 2000) early knowledge of verbs and verb frames is extremely idiosyncratic for each specific verb. In addition, morphological markings are unevenly distributed across verbs. In this view *I-am-eat-ing* is first learnt as an unanalyzed chunk and it takes the child a critical mass of verbs to realize that the frame *am—ing* can be used productively with different verbs. Two- and three-year olds have been found to generalize minimally, their repertoire consisting of a high number of conservative utterances and a low number of productive ones. The speculation is that a critical number of exemplars is vital for triggering schematization. Perhaps then, young children exploit n-gram statistics as a default option, because their knowledge of language is limited to a few *type* items. This situation is similar to learning in small set sizes and it only works if each string is learnt as a separate item. When children's repertoire is variable enough (arguably at ages three to four), then switching to change versus non-change as a source of information becomes more relevant and helps the learner reduce uncertainty by detecting variant versus invariant structure. The fact that learners in the large set size discard the middle item could be interpreted as a form of generalisation for material in the middle item position. This hypothesis will be tested in chapter 5. At this stage the link between AGL results and language learning can only remain speculative, but invites to intringuing research for the immediate future.

Chapter 3

The Variability effect: A graded, associative account

Since Reber's early studies (e.g., Reber, 1967), Artificial Grammar Learning (AGL) research has provided a steady stream of evidence that infants and adults become sensitive, after necessarily limited and often purely incidental exposure to complex stimuli, to the deep structure contained in chained events such as strings of letters. In a typical AGL situation, participants are first exposed to numerous stimuli and asked to memorize or process them in some way. Next, they are informed of the fact that the stimuli all instantiate a specific set of rules (a grammar), and asked to classify further strings as grammatical or not. Typically, participants can achieve some success in this classification task despite the fact that their verbalizable knowledge about the features that define grammaticality remains very limited. The learning mechanisms involved in such situations remain controversial. Recent results point to an inbuilt sensitivity to the transitional probabilities of adjacent items (Saffran, Aslin, & Newport, 1996). Other studies suggest fragment-based models involving memorised chunks of bigrams and trigrams (Dulany *et al.*, 1984; Perruchet & Pacteau, 1990; Servan-Schreiber & Anderson, 1990), learning of whole items (Vokey & Brooks, 1992), or learning based on similarity with previous items (Pothos & Bailey, 2000). Yet others postulate abstract learning of a distinct algebraic type in transfer tasks where the surface form of test items bears no resemblance to the training items (Reber, 1967; Marcus, Vijayan, Rao & Vishton, 1999).

The difficulty of identifying a single mechanism responsible for performance in AGL tasks should perhaps be taken as an indication that no such unique mechanism actually exists. Two points are worth highlighting in this respect. First, many of the proposed mechanisms actually turn out to be

equivalent at some level of description (Redington & Chater, 1998). Second, it appears likely that several sources of information might be used concurrently by subjects (as suggested by studies involving speech-like stimuli, e.g., Onnis, Monaghan, Chater, & Richmond, submitted).

The recent results by Gómez (2002), however, challenge virtually all extant AGL models. Gómez found that nonadjacent dependencies, that is, items that are structurally dependent but separated sequentially by one or more items, are learned better when the variability of the intervening items is large. In chapter 2 I have further found that nonadjacent dependencies were also learned better when the variability of the intervening items is zero (i.e., when there is only one possible intervening item). In other words, learning is best either when there are many possible intervening items or when there is just one such item, with degraded performance for conditions of intermediate variability. This U-shaped relationship between variability and performance cannot be readily explained by any of the putative mechanisms listed above. In particular simple associative mechanisms that rely on knowledge of chunks of items (or n-grams) would not predict such results, which thus appear to be incongruent with recent findings that both infants and adults can discover patterns in sequences based solely on sensitivity to low-level statistics (e.g. Saffran *et al.*, 1996). Gómez suggested that while humans are indeed attuned to distributional properties of the environment, they may also learn about which source of information is most likely to be useful, and that success might therefore depend specifically on the statistical properties of the stimulus environment they are exposed to. Crucially, Gómez's hypothesis is that learners capitalise on the most statistically reliable source of information in an attempt to reduce uncertainty (Gómez, 2003; Gibson, 1991;

Oaksford, & Chater, 1994; Chater, 1996). Thus, whether one becomes sensitive to the information contained in bigrams, trigrams or in nonadjacent structures may simply depend on the statistical properties of the specific environment that is being sampled.

The results obtained by Gómez and the ones charted in chapter 2 suggest that distributional learning is more powerful, dynamic, and data-driven than heretofore acknowledged, thus challenging the current fragment-based models. In this chapter, I aim to demonstrate that Simple Recurrent Networks (henceforth SRNs, see Elman, 1990; Cleeremans *et al.*, 1991) provide a unifying model that accounts for the dynamic U-shape pattern obtained experimentally. I discuss how connectionist networks can be seen as reducing uncertainty in a rational way (McClelland, 1998, Anderson, 1990). Performance strictly depends on developing separate graded internal representations of the hidden units for different nonadjacent dependencies in conditions of nil or high variability. It is suggested that reduction of uncertainty needs not be a conscious process, and might involve distribution of processing activity in a neural network. Perhaps then, humans are naturally and implicitly biased toward optimal learning.

My main goal in this chapter is to demonstrate that associative learning mechanisms are in fact sufficient to account for the u-shaped relationship between variability of the embedded material and learnability of the nonadjacent dependencies. However, and this is a crucially important point, not just any associative learning mechanism will do. In particular, I will suggest that successful learning of such material critically depends on the availability of graded, distributed representational systems, such as instantiated by connectionist networks. The graded character of the representations learned make

it possible for each such representation to simultaneously convey information about both the embedded material and the outer elements in such a way that learning can be "focused" (yet not through the action of any attention mechanism) on the most relevant source of information depending on the distributional properties of the entire material.

In the next subsection I present a connectionist simulation of Gómez (2002) for learning in highly variable contexts. Subsequently, I present a simulation of the results obtained in chapter 2 extending Gómez' data and incorporating learning with zero variability, resulting in a U-shape learning curve.

Simulation 1 – The Variability Effect Hypothesis

To summarise, Gómez exposed infants and adults to sentences of an artificial language of the form $A_i\_X_j\_B_i$, where $i \in \{1,2,3\}$. The language contained three families of nonadjacent pairs, notably $A_1\_\_B_1$, $A_2\_\_B_2$, and $A_3\_\_B_3$. Gómez manipulated the set-size of the middle element X_j in four conditions by systematically increasing the number from 2 to 6 to 12 and 24 word-like elements. In this way, conditional bigram probabilities $P(X_j|A_i)$ and trigram probabilities $P(B_i|A_i\_X_j)$ decreased as a function of number of middle words. In the test phase, participants were required to subtly discriminate correct nonadjacent dependencies, (e.g. $A_2\_X_1\_B_2$) from incorrect ones ($*A_2\_X_1\_B_1$). Notice that the incorrect sentences were new as trigrams, although both single words and bigram words ($A_2\_X_1$, $X_1\_B_2$, $X_1\_B_1$) had appeared in the training phase with the same frequencies. Hence the test required very fine distinctions to be made. Gómez hypothesised that if learners were focusing on n-gram

dependencies they should learn nonadjacent dependencies better when exposed to small sets of middle items because transitional probabilities between adjacent elements are higher for smaller than for larger set-sizes. Conversely, if learners spotted the invariant structure better in the larger set-size, Gómez hypothesised that increasing variability in the context must have led them to disregard the highly variable middle elements. Her results support the latter hypothesis: learners performed poorly with low variability whereas they were particularly good when the set-size of the middle item was largest (24 middle items).

Such scenario is problematic for associative learning mechanisms focused on processing local transition probabilities (i.e. from one element to the next) precisely because the embedded material appears to be wholly irrelevant to mastering the nonadjacencies: not only is there an infinite number of possible relative clauses that might separate *The dog* from *is*, but also structurally different nonadjacent dependencies might share the very same embedded material, as in *The dog that chased the cats is playful versus The dogs that chased the cats are playful* (Servan-Schreiber, Cleeremans, & McClelland, 1991).

While this state of affairs might suggest that nonadjacencies can only be mastered by structured, classical learning mechanisms (such as push-down automata), some authors have nevertheless suggested that associative learning mechanisms might in fact turn out to be sufficient to the extent that it is in fact seldom the case that the embedded material is completely independent from the head in natural language. To see this, consider for instance that a single dog and a pack of dogs are likely to be chasing different things. More generally, some embeddings are only possible after a singular dog (e.g., *The dog that scratched*

itself is very playful is grammatical, but **The dogs that scratched itself are very playful* is not) and others are only possible after a plural dogs (e.g., *The dogs that chased each other are very playful* vs. **The dog that chased each other is very playful*). Servan-Schreiber *et al.* (1991) demonstrated that associative learning mechanisms instantiated in Elman's SRN are sufficient to master such cases as long as the entire distribution of possible embeddings is statistically dependent on the head. This suggests that distributional approaches to language learning are more powerful than previously anticipated, provided that the environment contains even weak, statistical, relationships between the class of items that can form the elements of nonadjacent dependencies and the class of items that can form the embeddings.

To find out whether associative learning mechanisms can explain the variability effect, I trained an SRN (Elman, 1990; see Figure 3) to predict each element of sequences that were structurally identical to Gómez's material.

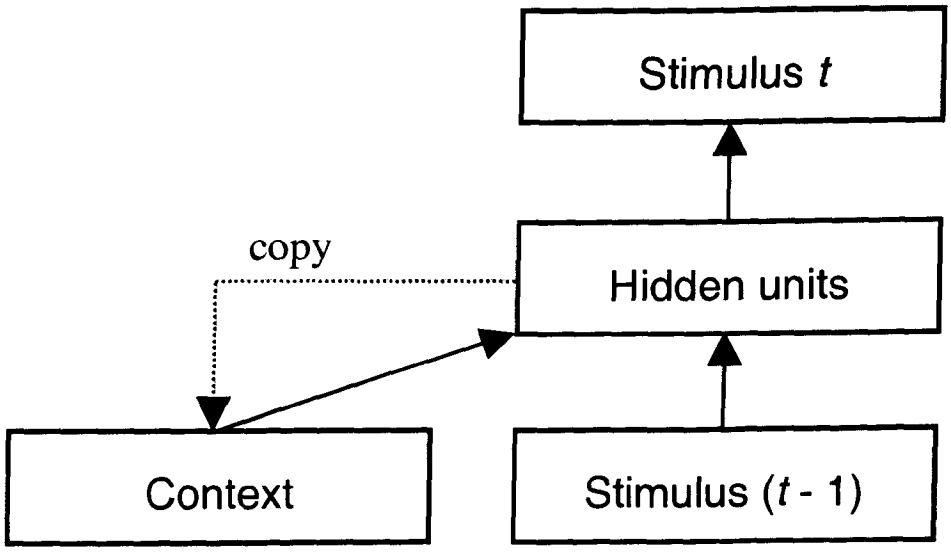

Figure 3. A Simple Recurrent Network (after Elman, 1990)

Method

Networks

48 SRNs [2] with different random initial weights between +.5 and -.5 were trained and tested, corresponding to 12 subjects in each of the four conditions in Gómez (2002). Single words were instantiated in localist representations[3], plus an End of Sentence (EOS) marker. Hence the networks had 31 input/output units (3 first words, 3 last words, and 24 potential middle words, 1 EOS), 15 hidden units, and 15 context units[4].

Materials

Training and test stimuli consisted of the strings generated from the finite-state grammar used by Gómez[5]. Test consisted of 6 grammatical and 6 ungrammatical strings.

Procedure

48 SRNs were trained and tested on a *prediction* task. At each stimulus presentation the task was to predict the next element in the string. After presentation of each string, the context unit activations were reset so that no information about the previous string was carried over. This corresponds to the networks receiving each string separately, as in the experimental setting. All networks in all conditions were given the same learning rate of .3 and

[2] All simulations implemented with the PDP simulator for Macintosh.
[3] Each word was an input vector with all units set to zero and a specific unit set to 1.
[4] Context units provided the network with a memory of previous instances by copying the activation of hidden units at time *t-1* and merging them with the activation of hidden units at time *t*.
[5] Gómez used two languages where the end-items were cross-balanced to control for potential confounds. Because all word vectors are orthogonal to each other, we created only 1 language.

Momentum of .9. Training consisted of the same overall number of token strings in all conditions (432 strings Gómez, 1080 strings for the networks). Weight update was carried out at the end of each string presentation[6]. The networks used the backpropagation learning algorithm and sum squared error as a measure of error.

Results and Analyses

To obtain the closest possible data contact with the experimental paradigm, I considered each network as a single participant and averaged together results from 12 separate networks in each condition. Being interested in the specific prediction of the third element (B_1, B_2, B_3) in each string, performance during test was assessed by recording the relative strength of the output unit corresponding to the correct successor of each middle element X_j. As a measure of weighted accuracy I used the Luce ratio (Luce, 1963), whereby the activation of the target output unit is divided by the sum of the activations of all output units[7]. A high Luce ratio indicates that most activation is placed on the correct target output unit, hence it can be taken as a measure of network's predicting power. To assess performance in a way that would correspond to human grammaticality judgement I computed the probability that each string would be classified as grammatical by entering the Luce ratio of the target output in the following standard expression (Dienes, 1992):

$$p(\text{"}grammatical\text{"}) = \frac{1}{1 + e^{-k \cdot Luce - T}}$$

[6] This is known as "online learning" as opposed to "batch learning", which consists of updating the weights after a number of training strings.

[7] Including target output, because in the case of the target unit "firing" alone, the weighted activation would be divided by zero.

where k is a scaling parameter and T is a threshold; both were adjusted manually so as to yield slightly higher numbers of grammatical than ungrammatical responses[8]. Probabilities over .5 were considered as a grammatical response while probabilities under .5 were considered as an ungrammatical response. The resulting probabilities for each test string were then averaged for each set condition over grammatical and ungrammatical sentences to yield global endorsement rates for each condition. Finally, I computed the percentages of correct classifications expected for grammatical and ungrammatical strings in each condition. The formula for correctness is:

$$correct = [p(G) - (1 - p(U))]/2$$

where p(G) is the probability of grammatical endorsement and p(U) is the probability of ungrammatical endorsement. It is these final values that I compared directly with Gómez' Experiment 1 results (see Figure 1 in chapter 2 and Figure 4 below). The SRN, trained in the same conditions as human subjects, displays a similar pattern of results. Performance does not improve gradually as a function of variability until a major boost occurs in condition 24. An analysis of the network's behaviour will clarify how and why performance increases with large variability only.

[8] This "positive bias" is a general trend in experimental settings where participants tend to respond YES more often than NO. Hence, k and T are not entirely free parameters, as their values are constrained by the requirement to reproduce the same overall positive bias found in the human experiments.

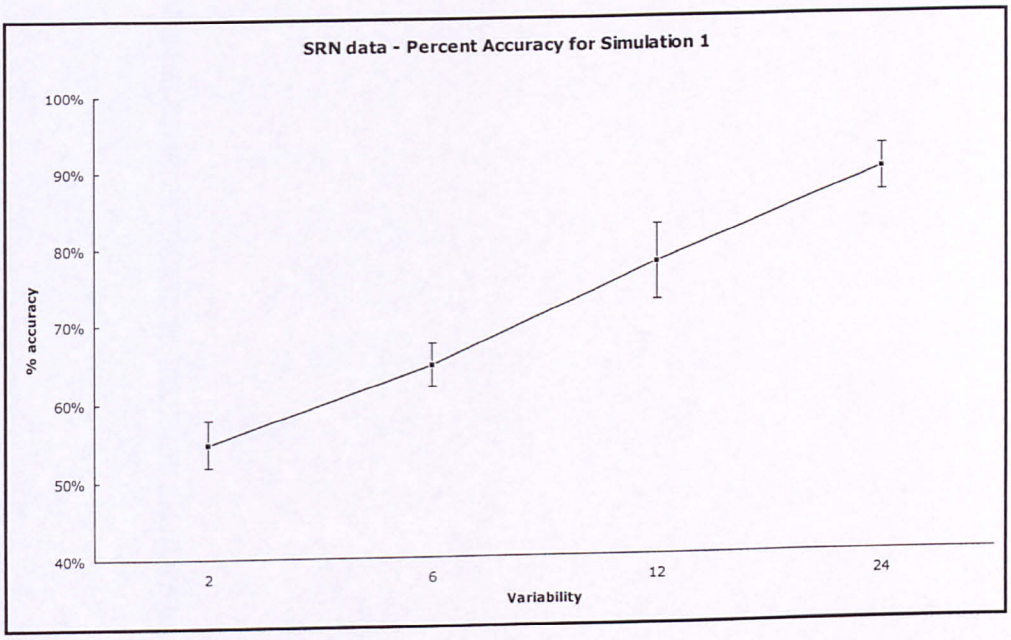

Figure 4. Percent accuracy for Simulation 1 across 4 conditions of variability.

Simulation 2 - The zero-variability hypothesis

Gómez attributed poor results in the middle set-sizes to low variability: the *variability effect* seems to be reliably effective only in the presence of a critical mass of middle items. In chapter 2 a new condition was further investigated when variability in middle position is eliminated, thus making the nonadjacent items variable. I replicated Gómez' experiment with adults and added a new condition, namely the zero-variability condition, in which there is only one middle element (e.g. $A_3\_X_1\_B_3$, $A_1\_X_1\_B_1$). They predicted that non-variability of the middle item would make the end-items stand out again, and hence detecting the appropriate nonadjacent relationships would become easier, increasing mean performance rates. Intuitively, sampling transitional probabilities with large context variability results in low information gain as the data are too few to be reliable; by the same token, the lack of variability should produce low information gain for transitional probabilities as well, because the probability $P(X_j|A_i)=1$, i.e. having seen any A_i will predict one X automatically. In other words, if learners try to reduce uncertainty they will ignore relations that just do not vary at all. Hence this should make nonadjacent dependencies stand out, as potentially more informative sources of information, by contrast. They obtained the final predicted picture of a U-shape learning curve (see Figure 2 in Chapter 2).

Method

Networks

60 SRNs were trained and tested, corresponding to 12 subjects in each of the five conditions in Chapter 2. The networks displayed the same structure and initialisation parameters as those in Simulation 1 above.

Materials

Training stimuli consisted of the strings generated from the finite-state grammar used by Gómez plus the new condition introduced in Chapter 2. Test stimuli consisted of 3 grammatical strings and 3 ungrammatical strings repeated twice, as in Chapter 2[9].

Procedure

The networks were trained in exactly the same way as in Simulation 1 above.

Results and Analyses

The results obtained are plotted in Figure 6 below. A U-shape similar to human data shows considerably better performance at end-point conditions.

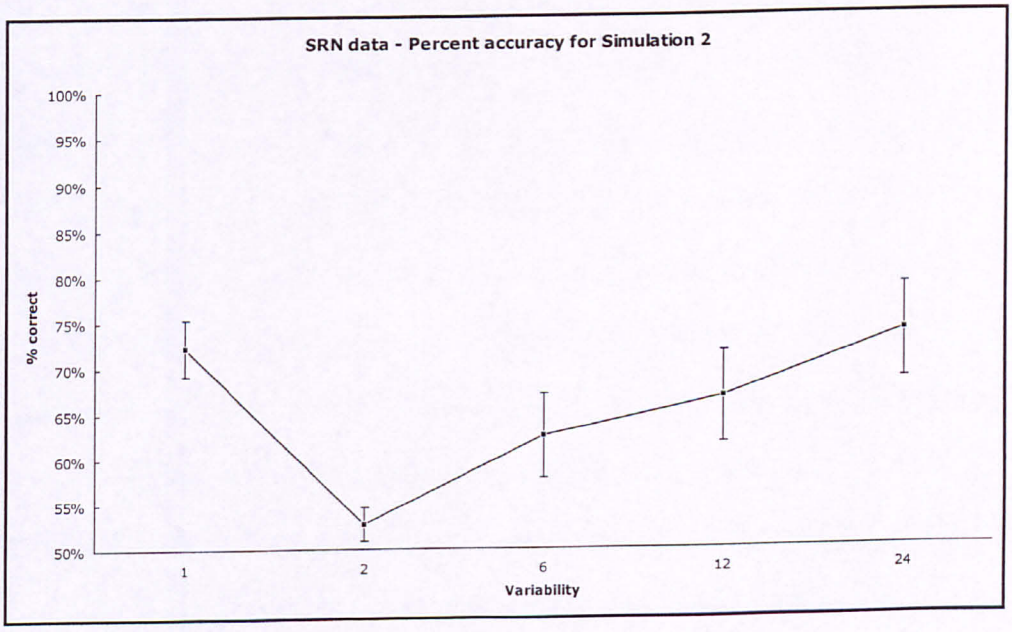

Figure 5. U-shape learning curve in SRNs for Simulation 2. Error bars are SEM.

[9] Given that in the new set-size 1 humans and networks are trained on one middle item they can only tested on strings containing one middle item, unlike Simulation 1. Hence networks were tested on 6 strings repeated twice. For this reason we ran new networks in simulation 2 even in set-sizes 2, 6, 12, and 24.

Learning nonadjacent structure in SRNs

An SRN trained to predict each element of sequences identical to those used by Gómez (2002) and in chapter 2 masters nonadjacencies in a manner that depends on the variability of the intervening material, thus replicating the empirically observed U-shaped relationship between variability and classification performance. In this section I provide an account of how the network learns about this material.

From the network's perspective the task, on each trial, is to predict the successor to the element it is presented with as input. This is difficult when the relevant information is contained not in the current input, but in previously experienced sequence elements, just as in Gómez. In such cases indeed, to achieve correct predictions of the tail element of a nonadjacency that spans irrelevant material, the network necessarily has to develop distinct internal representations in spite of identical inputs. This is easy if one imagines that a separate stream of processing can be dedicated to processing the embedding while maintaining information about the head, as traditional parsers such as push-down automata suggest. In the SRN, however, the internal representations associated with successive items are not stored or processed separately from each other, but rather they overlap in time. Achieving the required separation between internal representations in the face of identical inputs depends on the statistical properties of the input sequences and on the SRN's own architectural limitations. The graded character of the SRN's representational system prompted Servan-Schreiber *et al.* (1991) to describe such networks as *graded state machines.* Servan-Schreiber *et al.* showed that under certain conditions, these graded representations allow the processing of embedded material *and* of the material

that comes after the embedding, without duplicating the representations for intervening material. Hence the internal states of the SRN can be used *simultaneously* to indicate where the network is inside an embedding and, also to indicate the history of processing prior to the embedding. The identity of the initial element therefore simply *shades* the representation of states inside the embedding, so that corresponding elements have similar representations, and are processed using overlapping portions of the knowledge encoded in the connection weights. Yet the shading that the initial element provides carries information about the early part of the string through the embedding, thereby allowing the network to become sensitive to nonadjacent structure.

How does the network achieve this necessary separation of internal representations, and why does this process depend on the variability of the intervening material? It is useful to conceptualize learning in this situation as involving two opposite forces shaping the internal representations that the network develops over its hidden units: a top-down, error-dependent, force to produce the correct output, and a bottom-up, similarity-based, force to develop similar internal representations for similar sequences of elements (Servan-Schreiber, Cleeremans, & McClelland, 1991). Concerning the first factor, SRNs tend to converge towards the optimal conditional probabilities of observing a particular successive item to the sequence presented up to that point by minimizing the sum squared error (McClelland, 1998; Servan-Schreiber, Cleeremans, & McClelland, 1991). This amounts to exploiting first- and second-order conditionals. Unfortunately for our task, matching conditional probabilities only yields suboptimal behaviour. For first-order conditionals, when a head item A_i is presented, hidden units possess similar representations for predicting the

embedding X_j, because all instantiations of the embedding occur after A_i. This is also the case for predicting B_i, because any B_i occurs after any X_j. Backpropagation will also reduce the error by converging on the second-order conditional probability $P(B_i|A_i\_X_j)$. Interestingly, in Set-size 2 this probability is not extremely low (0.165) so we would expect the network to develop 6 separate representations, one for each string type. This process is overridden by the similarity of string types, which share the same embeddings. To discover the underlying structure the network has to "realise" that first- and second-order conditional probabilities lead to suboptimal solutions.

Information contained in context units acts as a concurrent force on hidden units from the bottom, helping maintain relevant non-local context information of previously seen items. Upon receiving an X_j, the context units preserve information about the previous item A_i. Backpropagation adjusts weights so that similar patterns on the output tend to be associated with similar patterns on the hidden units. In smaller set-size conditions similar representations develop for different tail predictions because the contribution from the shared embeddings is stronger than the contribution from the heads. To visualise the task, imagine that a trace of activation or *shading* from any head A_i has to filter through the flack of irrelevant embeddings. The trace carried over by each head item through the context units has to be strong enough to allow three different hidden unit representations, one for each nonadjacency. This trace activation competes with the shading contributed by the shared embedding. It increases as the strength of the embedding decreases, i.e. as the activation from embeddings on large set-sizes contributes less and less to shading the hidden unit patterns. With little variability, e.g. 2 *X*s, hidden units develop overlapping representations

for X_1 and X_2. Figure 6 presents the two principal components of a Multiple Dimensional Scaling (MDS) analysis over hidden units in condition 2, at the time of predicting the tail item over 15 different points in training. (Ungrammatical sequences are removed from the graph, because each produces exactly the same vector over the network's hidden units. Hence the graph displays 6 trajectories: one each for A_X_1, A_X_2, B_X_1, B_X_2, C_X_1, and C_X_2).

For successful separation of Xs and correct prediction of the successor to that X, the trajectories are expected to be cleanly separated at the end of training. By the same token, the trajectories corresponding to different Xs (X_1 vs. X_2) should be close to each other. Hidden unit trajectories move across training in the reduced 2-dimensional space, but they do not separate at the end of training. Contrast this result with Figure 7, a similar MDS analysis over the hidden units of a network in Set-size 24. Hidden units move together in space up to a point when they separate in 3 different regions of the 2-dimensional space, corresponding to 3 separate representations for A_1, A_2, and A_3. Hence, the presence of a large flack of 24 embeddings allows the trace from the relevant head item A_i to be maintained more strongly in the context units shaping the activation pattern of hidden units. In general, the networks are better able to preserve information about the predecessor of the embedded sequence across identical embeddings provided the ensemble of potential pathways is differentiated during training (Servan-Schreiber, Cleeremans, & McClelland, 1991). This is exactly the Variability Effect observed in human experiments.

Regarding the striking difference in performance between set-size 1 and 2, how do SRNs learn to predict the correct output in the former but not in the latter case? With variability comprised between 2 and 24 the networks reduce

error by providing a compact representation of the hidden units that groups embeddings together. Hence the information provided by the embeddings constitutes some sort of reduction of uncertainty in the form of information gain, although it leads to a suboptimal solution. Conversely, with zero variability the information contributed from the single embedding is minimal, i.e. it contributes nothing to reducing the error, as it is always the only item occurring in middle position. Hence the trace contributed by each specific head item suddenly stands out and becomes relevant enough to allow for separate hidden unit representations at the time of predicting the tail item. Strikingly, then, and somewhat counter-intuitively, learning in Set-size 1 and Set-size 24 seems guided by the same underlying principle. An MDS analysis of hidden unit trajectories (Figure 8) reveals that the network's behaviour is similar to the Set-size 24 condition: different trajectories are traversed ending in three distinct regions of the space.

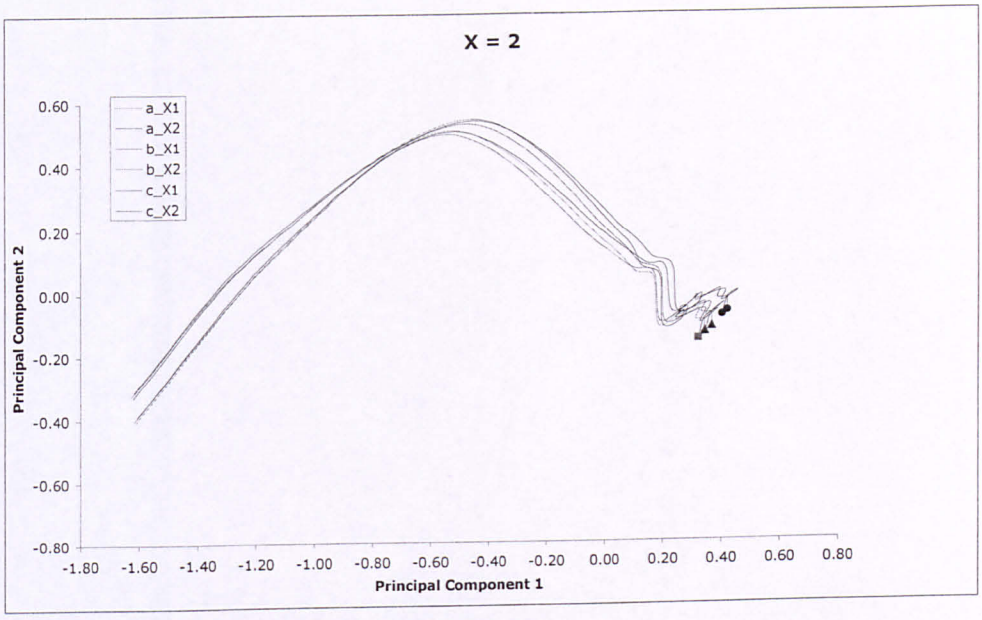

Figure 6. MDS analysis of hidden unit trajectories. A network trained on 2 *X*s fails to achieve the needed separation: all 6 trajectories remain close to each other all the way through the end of training. Hence the network can never form correct predictions of the successor to the *X*.

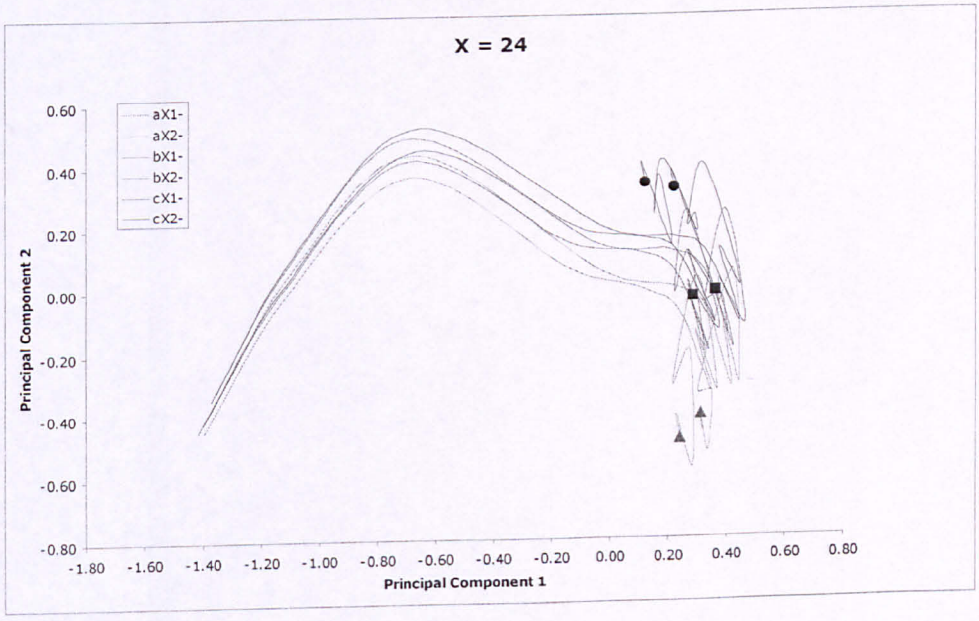

Figure 7. MDS analysis of hidden unit trajectories in the 24X condition: all 6 trajectories start out, on the left side, from the same small region, and progressively diverge to result in three pairs of two representations.

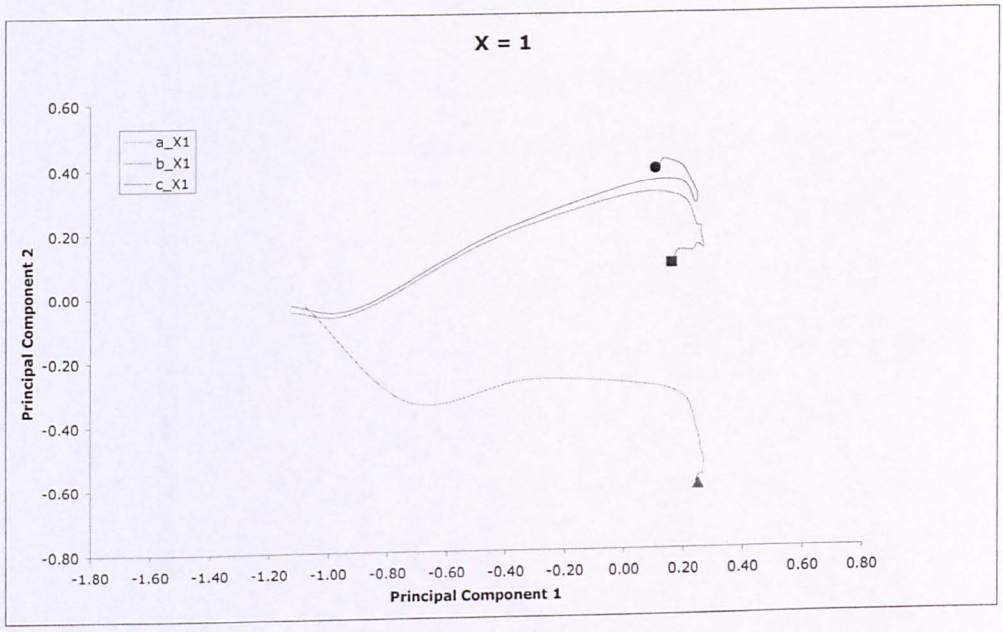

Figure 8. MDS analysis for a network trained on 1 X. Like in the 24X case, the network is successful in separating out the corresponding internal representations: The terminal points of each trajectory end up in different regions of space.

Conclusions

Sensitivity to transitional probabilities of various orders including nonadjacent probabilities in sequential learning has been observed experimentally in adults and children, suggesting that learners exploit these statistical properties of the input to detect structure. Detecting nonadjacent structure is central to learning natural languages and poses a genuine problem for simple associative models based on knowledge of adjacent items. Following Gómez (2002), a more elaborate proposal is that human learners may exploit different sources of information, including nonadjacent items, to reduce uncertainty. The amount of information gain provided by any element in the input may vary dramatically according to the statistical and informational landscape of the specific input.

In this chapter I have shown that SRNs succeed in accounting for the experimental U-shape patterns. This is not an easy feat, as in SRNs better predictions tend to converge towards the optimal conditional probabilities of observing a particular successor to the sequence presented up to that point. This means that minima are located at those points in weight space where the activations equal the optimal conditional probability. In fact, activations of outputs units corresponding to the three end items to be predicted in set-size 2, 6, and 12 settle around .33, which is the optimal conditional probability for $(B|X)$ across conditions. However, n-gram transitional probabilities may lead to suboptimal solutions, e.g. they fail to account for nonadjacent structural constraints. The network's ability to predict a nonadjacent element is modulated by variability of the intervening element, under conditions of either nil or high variability. This is achieved by developing separate graded representations in the hidden units. An analysis of hidden unit trajectories over training suggests that

the network's success at the end-points of the U curve might be supported by a similar type of learning, thus ruling out a simplistic rote learning explanation for Set-size 1. Together, the experimental and simulation data on the U curve challenge previous AGL accounts based on one default source of learning. Surprisingly, rather than ruling out associative mechanisms they suggest that statistical learning based on distributional information can be more powerful than heretofore acknowledged and dynamically attuned to the probabilistic properties of the environment.

Chapter 4

The Variability effect across modalities

Acquiring sequential information is vital in most domains of our life, from speech comprehension and production, to reading, to processing visual scenes, planning motor behaviour and action planning. To be able to detect sequential structures in the world at large humans need to exploit different sensory modalities, for instance auditory, visual, and tactile senses. A theoretically plausible hypothesis is that statistical learning is subserved by a single, domain-general cognitive mechanism, as Kirkham, Slemmer, & Johnson (2002) have suggested. Under this scenario, whether experimental stimuli are presented visually, auditorily, via tactile sensitivity, or whether different stimuli altogether are presented such as tones as opposed to syllables, similar performance effects are expected to transfer almost invariably across such conditions. Methodologically, gathering converging experimental evidence from different variations of the same experiment lends substantial robustness to the putative mechanism(s) under scrutiny. For instance, sensitivity to transitional probabilities of bigrams as evidenced by Saffran, Aslin, & Newport (1996) using chains of nonsense syllables has been corroborated by cross-modality studies. The same mechanism seems at work for example, in the segmentation of streams of tones (Saffran, Johnson, Aslin, & Newport, 1999) and in the temporal presentation of visual shapes (Fiser & Aslin, 2002). Such results support the view of a simple general-purpose and general-domain statistical mechanism sensitive to transitional probabilities in the input of any type.

My preliminary results in Chapter 2 on the adaptiveness of learning to different statistical landscapes and the emphasis on the potential cascade of cues

available to the learner (see Chapter 6) make it likely that modality constraints or modality preferences lead to different learning curves. A reduction of uncertainty hypothesis envisages a priori the possibility that perceptually salient modality-specific cues present in the input may drive structure building in modality-specific ways. In Chapter 6, for instance, I will show that adults' preference for certain plosive sounds in word-initial position is in itself *sufficient* to guide learner's choice for words versus part-words in a segmentation task, regardless of the structure underlying the words, and indeed even when no underlying structure is present. Further on, I will also show that generalisation to strings of the *A_X_B* type (see Chapter 2) containing a novel *X* item can be *sufficiently* explained away by mere preference for any word-initial syllable (again not the underlying structure) when this is made perceptually salient by a small preceding pause.

Conway and Christiansen (2002a) have conducted a series of comparisons across sensory modalities involving visual, auditory, and tactile senses to investigate the extent to which the three modalities afford sequential learning acquisition. Using stimuli from a standard AGL experiment in Gómez and Gerken (1999) they found that all three modalities performed well over chance and over control groups, but that there were also significant differences between modalities, with the auditory scoring at 96%, the visual scoring at 86%, and the tactile scoring at 74%. Conway and Christiansen (2002b) found further data supporting the view that "statistical learning processes are affected by modality constraints: vision is biased toward processing spatial input whereas audition is biased toward temporal input". Under this interpretation, sequential structure learning involves multiple, modality-constrained processes, which may

be tied to different non-overalapping brain areas. However, Conway and Christiansen also acknowledge that their results could be due to differential discriminability or perceptability of items according to the specific sensory domain recruited. This view would not necessarily imply differentiation of brain resources, but would be simply due to perceptual salience of certain cues. The issue of salience will be taken up in more depth in Chapter 6.

In this chapter I am interested in a preliminary investigation of the variability effect across the visual domain. In Chapter 2 I presented data in support of the variability effect in detecting nonadjacent dependencies using auditory stimuli. Below I present three variations of the variability experiment with visual presentation of the stimuli. In the first experiment, dubbed Visual Sequential, whole strings appear sequentially one at a time on the screen and are interleaved by white screen. In the second experiment, dubbed Visual Temporal, individual words within the strings appear one after the other on the screen, with a blank screen appearing between strings but not between words. The third experiment, Visual Sequential Abridged, is a replica of the first experiment but with half training, to test the effect of frequency of exposure on detecting structure visually. There were no strong predictions on what the results should be in these experimental variations. As a null hypothesis, I hypothesised that the same U-shape phenomenon found in the Auditory experiment in Chapter 2 should transfer across the board in the Visual experiments. The results are in fact more complex, and will be discussed along the way as well as in the general discussion below.

Experiment 2 - Visual Sequential (VS) version

Method

Participants

Sixty undergraduate and postgraduate students at the University of Warwick participated and were paid £3 each. None of them had participated in previous experiments.

Materials

The stimuli were identical to those used in Experiment 1, except that they were presented visually, in written form on a computer screen instead of auditorily.

Procedure

Exactly the same procedure as in Experiment 1 was used. Participants sat and looked at the strings as they appeared on the screen. Training lasted approximately 18 minutes, as in Experiment 1. Each string from the language was flashed up in black typeface against white background on a computer screen. Each string stayed on the screen for 2 seconds and was followed by a 750-ms white screen so that the strings could be perceived as independent one from the other. These values were chosen so that training lasted as long as training in Experiment 1. The test phase was the same as in Experiment 1, except that test stimuli were presented visually on the screen.

Results and discussion

An analysis of variance with Set Size (1 vs. 2 vs. 6 vs. 12 vs. 24) and Materials (L1 vs. L2) as between-subjects and Grammaticality (trained vs. untrained strings) as a within-subjects variable resulted in a main effect of Grammaticality, $F(1,50) = 16.39$, $p < .001$, but no other significant main effects or interactions. Comparisons between adjacent set-size conditions revealed no significant differences, in particular no significant increase in performance between set size 12 and set size 24, $t(22) = 1.395$, $p = .177$, nor a significant decrease in performance between set size 1 and set size 2, $t(22) = 1.697$, $p = .104$. In contrast to Experiment 1, a polynomial trend analysis did not show a significant quadratic effect, $F<1$. Figure 9 presents the percentage of endorsements for total accuracy in each of the five set-size conditions. Table 1 below presents the percentage of endorsements for trained versus untrained strings and total accuracy in each of the five set-size conditions.

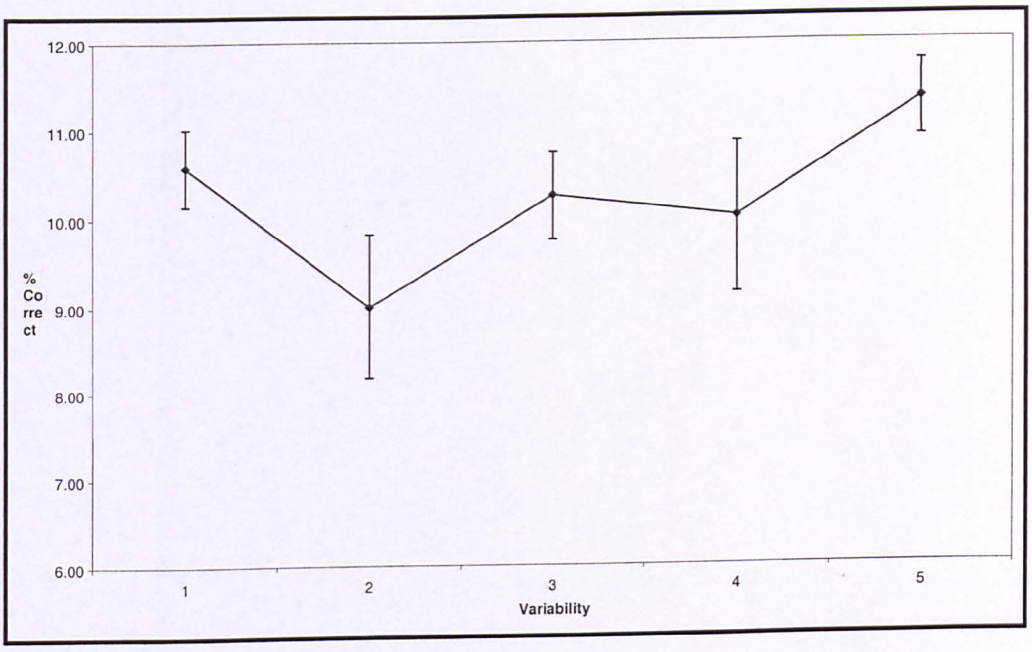

Figure 9. Total percentage endorsements in Experiment 2 for different variability.

Set Size	Trained	Untrained	Total
1	97%	79%	88%
2	90%	59%	75%
6	95%	75%	85%
12	90%	76%	83%
24	97%	91%	94%

Table 1. Percentage of endorsements for trained versus untrained strings and total accuracy in each of the five set-size conditions.

The confirmatory bias in the Variability Experiments

In order to compare directly the Auditory significant results and the Visual nonsignificant results this section discusses the effects of a positive bias on individuals' responses for both experiments. In particular, I pit the positive bias scores obtained in the Auditory version against the Visual Sequential version, because the latter comes closer to reproducing a U shape.

The positive bias is a well-known phenomenon in experimental psychology, whereby participants required to express judgements in the form of binary yes/no choices have a natural tendency to choose YES more often than NO, i.e. they tend to confirm rather than disconfirm an option. This seems to be a particular case of a more general case for the human tendency to seek out information that confirms our beliefs rather than looking for that which disproves it.

The confirmatory bias was found in both the Visual Sequential and Auditory versions of my experiment, despite participants being explicitly informed of the presence of an equal number of correct and incorrect test stimuli among the 12 test items. Methodologically, it might be contended that extreme values of positive bias in participants' responses introduce noise in the results, on the basis that some participants failed to understand the explicitly required test instructions. The Visual Sequential version displays the predicted variability effect only to an extent (there is no significant Grammaticality x Set Size interaction and *t*-tests between adjacent conditions - Set Size 1 and 2 and Set Size 12 and 24- are not significant). Hence we may want to evaluate whether the impact of the positive bias bears any relevance to the weakness of the findings. If a participant shows a particularly strong positive bias we may want to discard

his/her performance and run a new participant. It must be noted that a high positive bias correlates negatively with performance values (at least in the case of forced binary choices where the number of required "yes" responses are the same as "no" responses). In the extreme case of a response bias of 12 (i.e. all test items are responded to as "yes") performance drops at chance level, because only 6/12 test items are correct when responding "yes". In general, because the positive bias is a ubiquitous phenomenon in experimental settings it poses a problem only for extreme values (10-12 in our case). Tables 2 and 3 below show number of individuals in each condition in the Visual Sequential and Auditory experiments, for biases equal or higher than 9, 10, and 11 respectively.

Positive bias	SET SIZE 1	SET SIZE 2	SET SIZE 6	SET SIZE 12	SET SIZE 24	Total
≥ 11	1	4	1	1	0	7
≥ 10	1	4	1	1	0	7
≥ 9	2	4	2	1	1	10

Table 2. Positive bias for the Visual Sequential experiment

Positive bias	SET SIZE 1	SET SIZE 2	SET SIZE 6	SET SIZE 12	SET SIZE 24	Total
≥ 11	0	1	1	1	0	3
≥ 10	0	2	1	3	0	6
≥ 9	0	2	1	5	1	9

Table 3. Positive bias for the Auditory experiment.

From a comparison of the totals in the two tables above we can see that Visual Sequential has twice the number of extreme positive bias values (i.e. ≥ 11) than the Auditory, though overall the same number of values that are ≥ 9. In addition, the distribution of such values tends to affect the middle points of the U-shape curve (conditions 2, 6, and 12), not the end-points, in both experiments. This does not really run counter to my expected results, as high positive bias values mean poor performance, as indeed predicted in these conditions. In fact one could interpret high positive bias results as a failure of participants to actually detect the correct structure. Bearing in mind that incorrect test stimuli are very similar to correct ones it is not surprising that bad performers tend to press "Correct" most of the times in the face of indecision and uncertainty.

I conclude that the positive bias is not affecting the current results in the Visual Sequential version. It mainly manifests itself as a biproduct of uncertainty in conditions of middle variability where I expect participants to be confused and score near chance levels. Because the distribution of positive bias values in the two experiments is similar it does not contribute to explaining why the Visual Sequential results are not significant. The bow in the Auditory version is more marked than in the Visual Sequential version. This could be interpreted as a ceiling effect in the Visual Sequential, due to overtraining. To test this hypothesis I ran the same Visual Sequential experiment with half training trials. The experiment is presented below.

Experiment 3 - Visual Sequential Abridged version (VSA)

The results from the Visual Sequential version above are not significant, although from Figure 9 one can see that there is a hint of a U-shape in the trend, but there is a higher proportion of participants who successfully detect the invariant structure in middle size conditions as well. In all variations of the variability experiment reported here so far the 432 training strings were organised in 3 sets, corresponding to 144 training strings per set, interrupted by a break section to allow for participants to take a short break. Perhaps the not-so-marked bow may be due to an overtraining artifact coupled with the fact that the visual display of whole strings may make the detection task easier for participants, thus "giving away" the underlying structure sooner. To test this hypothesis, I ran an "abridged" version of the Visual Sequential using half the training tokens, i.e. 216 training strings, in the hope that a shorter training would make the task slightly more difficult, thus compensating the putative facilitatory effect of the visual presentation.

Method

Participants

Fifty undergraduate and postgraduate students at the University of Warwick participated and were paid £3 each. None of them had participated in previous experiments.

Materials.

Training and test stimuli were identical to those used in the Visual Sequential Experiment.

Procedure.

Exactly the same procedure as in the VS Experiment was used, except that training consisted of 216 strings, exactly half as many as in the VS Experiment.

Results and discussion

An analysis of variance with Set Size (1 vs. 2 vs. 6 vs. 12 vs. 24) and Materials (L1 vs. L2) as between-subjects and Grammaticality (trained vs. untrained strings) as a within-subjects variable resulted in a main effect of Grammaticality, $F(1,40)= 5.199$, $p< .027$, but no other significant effect. Comparisons between adjacent set-size conditions revealed no significant differences, although a few were not distant from significance, namely between set size 6 and set size 12, $t(22)= 1.848$, $p= .078$ and between set size 12 and set size 24, $t(22)= 1.881$, $p= .073$. The results disconfirm that a shorter training for visual presentation should result a neater U-shape, indeed by looking at Figure 10 there is an almost inverted curve to the one predicted, with performance low for set size 1 increasing at a peak for set size 6, decreasing again at set size 12 and increasing back again for set size 24. Such patchy results are not entirely understood at present and suggest that other aspects of performance than mere token reduction may play a role in determining higher rates of successful detection of the underlying structure.

Because the visual versions above display sentences one at a time instead of one word at a time, they do not strictly reproduce the equivalent temporality of the auditory version in which each word was heard in sequence. For this reason I ran the following Visual Temporal experiment.

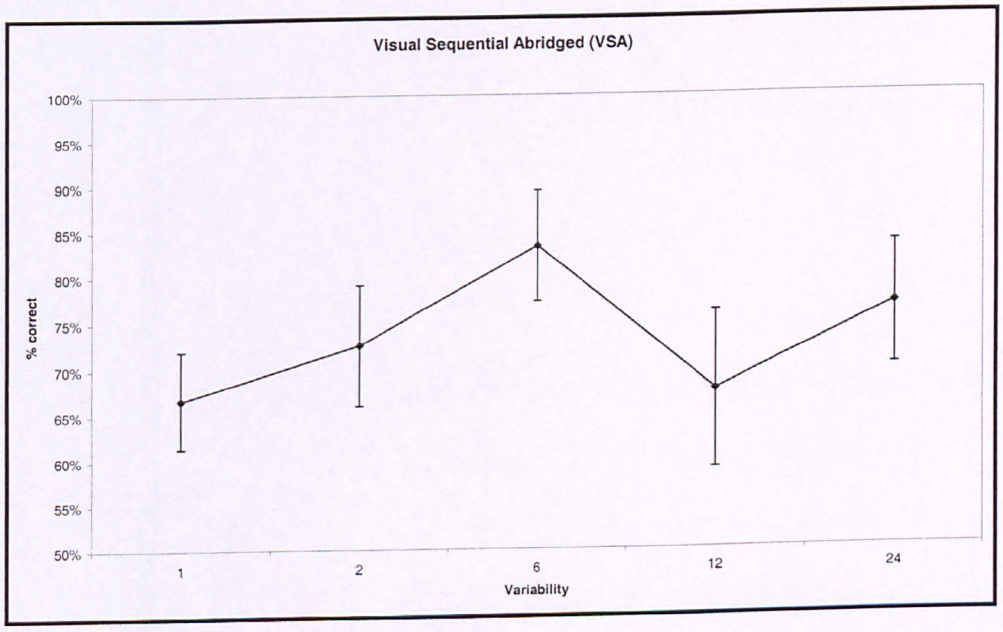

Figure 10. Percent correct responses for Experiment 3.

Experiment 4 - Visual Temporal (VT) version

Method

Participants

Sixty undergraduate and postgraduate students at the University of Warwick participated and were paid £3 each. None of them had participated in previous experiments.

Materials

The stimuli and presentation were identical to those used in Experiment 2. Visual presentation, however, differed in that individual words were presented one at a time.

Procedure

Exactly the same procedure as in Experiment ? was used. Participants sat and looked at the stimuli as they appeared on the screen. Training lasted approximately 18 minutes, as in Experiment 1. This time each word from the language was flashed up individually in black typeface against white background on a computer screen. Each word stayed on the screen for 666 ms and was immediately followed by the next word without a blank screen. Each end-of-string word was followed by a 750-ms white screen, so that the strings could be perceived as independent one from the other. These timings were chosen so that training lasted as long as training in Experiment 1 and 2. The test phase was the same as in Experiment 1, except that test stimuli were presented visually on the

screen and one word at a time (using the same timing as training stimuli) for congruity with the training phase.

Results and discussion

An analysis of variance with Set Size (1 vs. 2 vs. 6 vs. 12 vs. 24) and materials (L1 vs. L2) as between-subjects and Grammaticality (trained vs. untrained strings) as a within-subjects variable resulted in a main effect of Grammaticality, $F(1,50)= 5.199$, $p< .05$, but no other significant effect. Comparisons between adjacent set-size conditions revealed no significant differences, although a few were not distant from significance, namely between set size 6 and set size 12, $t(22)= 1.848$, $p= .078$ and between set size 12 and set size 24, $t(22)= 1.881$, $p= .073$. Figure 11 summarises the data. Table 2 below presents the percentage of endorsements for trained versus untrained strings and total accuracy in each of the five set-size conditions (Visual Temporal version).

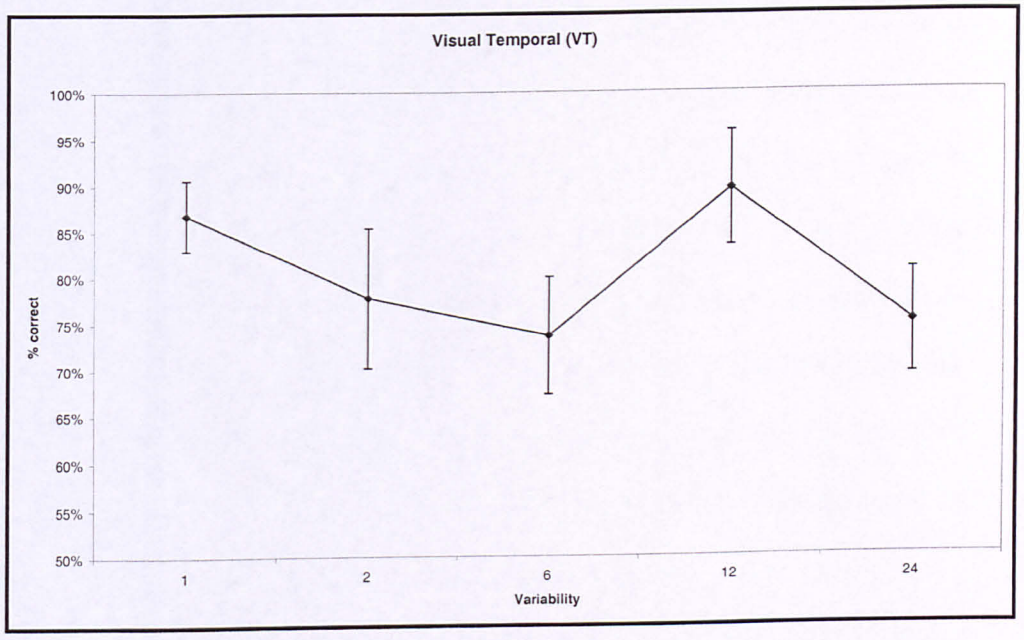

Figure 11. Percent correct responses for Experiment 4.

Set Size	Trained	Untrained	Total
1	91%	81%	86%
2	81%	73%	77%
6	80%	66%	73%
12	90%	88%	89%
24	77%	72%	75%

Table 4. Percent correct responses for Experiment 4 expressed in terms of seen (trained) and unseen (untrained) items recognised correctly.

General Discussion

As a null hypothesis, I started on the assumption that the same variability effect found in the original auditory experiment in Chapter 2 should transfer across perceptual domains, for instance visual presentation of training and test strings. In this chapter a battery of three visual variations were presented, namely VS, VT and VSA. For the VS it looks as if the variability effect is there, although with little statistical power. One plausible, intuitive explanation envisaged here is that the visual task is cognitively easier, and thus a ceiling counter-effect pushes the bow in midsize conditions up to levels of performance higher than predicted. In other words, detecting the relationship between the first and the last word in a string appears easier when the whole string is presented visually on the screen, regardless of middle item variability, perhaps because the two end-items are both present at any one time. Several participants reported that training was rather long. Given the successful results obtained by several AGL experimenters who tested participants for few minutes (e.g. Saffran *et al.*, 1996; Onnis, Monaghan, Chater, & Richmond, submitted; see also chapters 6 and 7), if not seconds (Jusczyk & Aslin, 1995), there was reason to believe that my participants were overtrained. Overtraining usually produces several confusing side-effects that are not directly separable and contribute to noise in the data. Firstly, overtraining boosts the effect of frequency of exposure to a given item, then promoting rote-learning to the detriment of structure building. Secondly, long training sessions diminish participants' attention dramatically, to the point that some participants may stop attending at the task. Once they re-attend to the task – later in the training or at test– they may have been distracted long enough to have

"cancelled" the effects of learning accumulated thus far. In both cases, this produces noise in the data.

Unfortunately, the VSA version did not yield the expected clear picture. In some sense, the variability effect is reversed, with lower performance in set size 1 and a peak of performance in set size 6. Such results are not easily interpretable at present and more experiments need to be carried out to extricate their significance.

Turning to the VT version (which presented words one at a time on the screen, with no blank screen between words of a sentence and a blank screen between sentences) an S-shape is displayed with peaks of performance at set sizes 1 and 12 and valleys at set sizes 2, 6, and 24. With the limited knowledge I have of the variability phenomenon, this curve is not intuitively straightforward to account for. Both the VS and VT versions have a spatial as well as temporal dimension because items appear on the screen at different times. However, whereas it is plausible to assume that outside experimental settings people happen to read one sentence at a time quite frequently as in the VS version (for instance, peeking at an advertisement in the street, or glancing at the title of a book), it is far less ecologically plausible that people read words in a sentence one at a time (some TV commercials exploit this technique to attract audience, but it is reasonable to assume that this happens quite infrequently). Hence, while the VT is the structurally direct analogue of the Auditory version, it is not probably the direct analogue in the real world. The decrement in set size 24, however, may have two different explanations entirely consistent with the variability hypothesis. Firstly, Gómez stressed in her original paper that a variability of 24 was not to be taken as an absolute value, rather that a critical

mass of variability is needed to detect invariant structure. A priori, then, this critical mass may be "shifting" in dimension across variations of the same task according to specific complexity. Hence it is perfectly possible that the critical mass for the VT version is around 12 rather than 24. If correct, this interpretation has to explain why performance drops again at set size 24. So far, a tacit assumption of the variability effect has been that beyond the critical mass learners should maintain high performance levels. Hence, in an ideal infinity hypothesis condition in which each string presented has a new middle item, performance should be at least as good as with the minimal critical mass required to identify invariant structure. The theoretical insight underpinning the infinity hypothesis is that learners are sensitive to change versus no change and the more change the better to trigger individuation of invariant structure. However, this hypothesis has not been tested here and it remains entirely plausible that beyond a given amount of variability (say 12 or 24) –let us call it the "specific variability" hypothesis - detecting structure should be easier or indeed more difficult. The decrease in set size 24 for the VT version indicates that it might be difficult in a visual temporal task to detect invariant structure with more than say 12 variant middle items, but I do not have data for larger set sizes than 24 in neither of my experiments. As a conclusion, I have just scratched the surface of a large project. The promising results obtained so far have highlighted the possibility of a critical mass or a specific mass of variability supporting detection of invariant structure. The specific mechanisms of interaction with different sensory modalities is not fully understood at present. The next chapters will further elucidate the possibility of detecting nonadjacent invariant structure in order to bootstrap speech segmentation and generalisation.

Chapter 5

Bootstrapping abstract linguistic representations

Throughout the cognitive sciences generalisation is seen as the hallmark of cognition. Understanding how humans generalise is thought to contribute a central piece of the puzzle to how the mind works. Given limited exposure to a set of stimuli, infants and adults are able to "go beyond the data" by building representations that are abstract at some level of analysis. In the study of the language faculty, in particular, generative linguistics (e.g. Chomsky, 1957) has highlighted the mind's extraordinary power in extracting abstract syntactic representations given limited and degenerate exposure to language samples. Two fundamental arguments at the heart of the generative program in linguistics have made a significant impact on the field of language acquisition:

a) children and adults produce in their life thousands of novel sentences that they have never heard before and for which knowledge of previously encountered sentences cannot account as sufficient knowledge. Hence the mind must come equipped with some forms of in-built innate language-specific and species-specific knowledge to guide learners. This is known as the poverty of the stimulus argument.

b) the core knowledge of language is mainly knowledge of syntactic rules. Such rules have long been described as propositional and algebraic in nature, although subsymbolic systems like connectionist networks have been shown to generalise to novel instances too. In addition, AGL experiments have been used at different times to show that generalisation can be supported using distributional and other potentially relevant cues present in the input.

While the argument from the poverty of the stimulus will be dealt with at length in chapters 7 and 8, and the rules-versus-statistics debate will be covered in chapter 6, this chapter focuses on the possibility that at least some of the abstract linguistic representations that have traditionally been ascribed to innate knowledge might in fact be bootstrapped from experience using distributional learning.

The words of natural languages are organised into categories such as NOUN, VERB, ARTICLE, etc. that form the building blocks for constructing sentences. Hence, a fundamental part of a language user is the ability to identify the category to which a specific word, say *apple*, belongs. The process by which language learners bootstrap lexical category membership is not fully understood. Some researchers (e.g. Pinker, 1984) have seen the problem as one of mapping between prior semantic categories such as object and action and the set of innately specified syntactic categories. This semantic boostrapping would make use of children's knowledge about word meanings as a basis for an initial classification of words. Others have proposed phonological contraints (e.g. Gleitman, Gleitman, Landau & Wanner, 1988) based on the fact that members of different word classes, e.g. nouns versus verbs display different phonological regularities. For instance, stress in English disyllabic nouns tend to fall on the initial syllable whereas in verbs it falls predominantly on final syllables, and English polysillabic words are mainly nouns (Cassidy & Kelly, 1991). Another form of information proposed for boostrapping word classes are prosodic cues such as the mutual predictability between the way a sentence is constructed and the way it is said, i.e. its prosodic phrasing (Morgan & Newport, 1981). All these are viable hypotheses.

The preferred proposal that I will follow here about how children go about grouping words into relevant categories is that they perform a distributional analysis on the sentences they hear and start categorising together words that appear in the same lexical co-occurrence patterns.

As mentioned in earlier in this thesis, one of the fiercest arguments levelled at distributional learning concerns the uninformativeness of such mechanisms for detecting linguistically relevant properties (Pinker, 1984). Among a series of criticisms Pinker argues that the properties of the raw input that can be detected using distributional learning pertain to serial position, adjacency, and cooccurrence relations among words, whereas "most linguistically relevant properties are abstract, pertaining to phrase structure configurations, syntactic categories, grammatical relations, [...] but these abstract properties are just the ones that the child cannot detect in the input prior to learning" (Pinker, 1984 p.49-50). Several scholars (e.g. Redington, Chater, & Finch, 1998) have counterargued that the utility of distributional statistics lies not in describing the relevant abstract linguistic properties but in helping the learning child to extract such abstract representations from the input. These studies have made successful use of computational and statistical analyses of child-directed speech in large corpora (Cartwright & Brent, 1997; Kiss, 1973; Mintz, Newport, & Bever, 1995, 2002; Redington, Chater, & Finch, 1998). Redington, Chater, & Finch (1998), for instance, have shown that highly local and extremely simple distributional statistics collected over large corpora of text such as the CHILDES database and the British National Corpus (BNC) are informative in discriminating nouns, verbs, adjectives and closed-class words with cluster analysis. This valuable work has

shown that distributional information is, in principle, a very useful cue for boostrapping syntactic categories, but it has yet to be demonstrated whether young children, and learners in general, practically can and do utilise this source of information. Promising results have been obtained using AGLs (Gómez & Gerken, 1999; Maratsos & Chalkley, 1980; Mintz, 2002) and the present work is meant to contribute new evidence that adults are able to build syntactic-like categories from the raw input they receive. Given the relevance of nonadjacent structure for language acquisition and the findings that its detection is modulated by the amount of variability of embedded material as highlighted in chapter 2, the aim of this chapter is to establish empirically whether detection of nonadjacent frames can support generalisations to new embeddings, via a process of categorisation of the class EMBEDDING. Below I provide the rationale for doing so and subsequently present a new experiment.

Generalisation under conditions of variability

As discussed above distributional information appears to be, in principle, a powerful source of information for discovering syntactic classes. However, investigations of the claim that learners actually perform a distributional analysis over instances of artificial grammars have met in the past with considerable experimental limitations. For instance, Smith (1966) was interested in whether learners were able to extract lexical co-occurrence patterns. He trained adults on an artificial grammar containing 4 non-overlapping categories *M*, *N*, *P*, and *Q* that were arranged in two types of sentences:

S -> MN

S -> PQ

Each category contained four words, and learners were trained on 24 of the 32 possible sentences. At test, they were asked to decide whether they had heard a sentence among the following:

a) heard sentences

b) unheard grammatical combinations, adhering to the *MN/PQ* pattern

c) sentences adhering to an *MQ/PN* pattern

d) ungrammatical sequences (e.g. *PM*, *QP*, etc.).

He was hoping to find that b) should be preferred to c) and c) should be preferred to d), but instead he found that both b) and c) responses were significantly greater than d). The results suggest that learners had generalised according to the absolute position of the words (first or second), but not their relative position based on the lexical co-occurrence patterns (for instance that words belonging to the *P* category only co-occur with words of the *Q* category). Further studies using the same paradigm as Smith found that generalisation of relative position was possible only in the presence of extra converging cues attached to some portion of the words, such as salient affixes (Braine, 1987; Frigo & McDonald, 1998; Gómez & Gerken, 1999). The converging cues seem to act as a necessary bootstrap into the distributional patterns that are relevant for category abstraction.

In a recent paper Mintz (2002) reasoned that the type of language used in Smith (1966) and in following studies might provide too limited distributional cues to engage distributional learning mechanisms, as all sentences were only 2 words long, whereas natural languages typically contain a richer distributional environment. He devised a language similar to Gómez (2002), with four shared static frames and four medial elements and found that category generalisation was supported in classifying medial words based on the surrounding frame. Mintz (2002) further elaborated that in performing a distributional analysis a word can be both a target word for categorisation while at the same time functioning as a categorising element. But while an ideal learner may entertain words as targets and environments simultaneously, actual learners may in fact need more reliable cues in order to consistently treat a word as either target *or* environment. He argued that this may not be possible using the two-word MN/PQ paradigm as there is no basis for making the above distinction. Conversely, the slot-and-frame grammar using 3 words might provide a grounding for distributional analysis by functioning as a figure-ground distinction. This is consistent with the line of argument followed in previous chapters that learners are sensitive to change versus non-change. Crucially, Mintz only provided general considerations about the figure/ground distinction and the role of frames in language acquisition (but see Mintz, in press), hence this work contributes a follow-up and more detailed account based on the variability effect. In particular, the question tackled here in more depth is as follows: is detection of frames in reference to embedded words (as found in chapters 2 and 3) a separate process from tracking the pattern of middle words in reference to frames? Or, alternatively, does detection of invariant

nonadjacencies afford generalisation of middle items as belonging to the same syntactic-like category? My hypothesis is that if the two processes are two sides of the same coin, generalisation to a new middle element in the experienced frames should occur only in conditions of nil or large variability of the middle item category.

Mintz also left open the extent to which sequences of words in natural languages actually display an alternation of frames of the type simulated in his language. In chapter 2 it was remarked that the asymmetry in the distribution of open class words and closed class words in natural languages such as English may effectively help learners detect syntactic constructions that sequentially span one or several words. Such nonadjacent dependencies are fundamental to the process of progressively building syntactic knowledge of, for instance, tense marking, singular and plural markings, etc. Crucially it has been proposed that these constructions may function as frames or "construction islands" (see Tomasello, 2003 for an overview) for subsequently building abstract and productive construction patterns. For instance, Childers & Tomasello (2001) tested the ability of 2 ½-year-old children to produce a verb-general transitive utterance with a nonce verb. They found that children were best at generalising if they had been mainly trained on the consistent pronoun frame *He's VERB-ing* (e.g. *He's kicking it, He's eating it*) rather than on several utterances containing unsystematic correlations between the agent and the patient slots (*Mary's kicking the ball, John's pushing the chair*, etc.).

The argument in this chapter is that while detection of such syntactic frames may be achieved under conditions of no variability or large variability by a

focus on what changes versus what stays invariant, thus leading to "discard" the common embeddings in some way, there may be a reversal and beneficial effect in noting that common elements all share the same contextual frames. It is reasonable to hypothesise, then, that if several words whose syntactic properties and category are unknown are shared by a number of contexts, then they will be more likely to be grouped under the same syntactic label, for instance VERB. Consider a child that is faced with discovering the class of words such as *break, drink, build.* As the words share the same contexts below, a learner may be driven to start extracting a representation of the VERB class:

am-X-ing

dont-X-it

Lets-X-now!

Most importantly, in hearing a new word in the same familiar contexts, for instance *eat* in *am-eat-ing,* the learner may be drawn to infer that the new word is a VERB. Ultimately, having categorised in such a way, the learner may extend the usage of *eat* as a VERB to new syntactic constructions in which instances of the category VERB typically occur. For instance s/he may produce a novel sentence *Lets-eat-now!* Applying a category label greatly enhances the generative power of a linguist system.

In continuing to test empirically the viability of category abstraction through distributional analysis of the input, the specific question that is being asked in this chapter is whether generalisation to new *X* items in the *A_X_B*

artificial grammar used in previous chapters is supported under the same conditions of zero or large variability that afford the detection of invariant structure. The specific prediction is that detection of invariant contextual structure and generalisation to new elements allowable within the invariant structure are two sides of the same coin. If constructional frames are acquired under the variability hypothesis, generalisation will be supported when there is no variability of middle elements as well as when there is large variability of middle elements. Likewise, because invariant structure detection is poor in conditions of middle variability, generalisation is expected to be equally poor in those conditions too.

As reiterated throughout this dissertation, it is fundamental to establish analogies as direct as possible between the artificial grammars constructed and the particular aspects of natural language that such grammars are meant to reproduce. In predicting a U shape for generalisation I want to motivate such results in the light of correlated data from the acquisition literature. The case for generalisation under large variability has an analogy in the literature under the "critical mass hypothesis" (Marchman & Bates, 1994). Children's early period of productive language seems characterised by a strong conservatism with regard to the utterances heard. Tomasello (2002, for a review) has noted that early use of verbs is restricted to contextual frames, or islands, for each specific verb. Children gradually become more productive with novel verbs and with known verbs in new contexts at the age of 3-4 years. Likewise, Pine & Lieven (1997) found that the articles "a" and "the" are used with different nouns at the age of 2-3 years, and Pizzuto & Caselli (1992) have found similar results for Italian morphology. This

suggests an item-based process of learning, where children originally possess no adult-like abstract knowledge of what constitutes a syntactic category like VERB or ARTICLE and construct such knowledge gradually from the items. The assumption is that there is a critical mass of exemplars of particular utterances necessary to trigger the process of categorisation. However, Tomasello points out that the specific nature of the critical mass remains vague. For instance, it fails to describe whether types or tokens have to reach the critical mass and the relation between types and tokens. The prediction of generalisation under large variability is in line with the critical mass hypothesis. Moreover, it helps specify the hypothesis further, in support of the crucial role of types rather than tokens for triggering categorisation. This is because in the experiment as variability increases the number of *type* sentences increases while the number of *token* sentences in each condition stays the same (432). Hence the relative type/token ratio is proportional to variability.

The prediction of generalisation under no variability is seducing because it allows a qualitative and counterintuitive prediction. Remember that in chapter 2 excellent performance under zero variability was *interpreted* as detection of invariant structure. However, another equally plausible interpretation remained open, namely that learners trivially memorise the three type sentences as trigrams during learning, given extended exposure to them (each occurs 144 times), and then discriminate them successfully against similar but unheard trigrams at test. This latter hypothesis trivialises the U shape outcome because it means that nonadjacent dependencies are not detected under zero variability and the fact that humans learn a few instances by heart is neither surprising nor informative for a

theory of learning. However, if learners were to show an ability to *generalise* to a novel embedding under no variability, this would suggest that they do not memorise sentences as trigrams, but rather extract the dependency between the head and tail of the strings, and sanction as grammatical a novel sentence that contains a novel embedding, but not a novel sentence that contains a novel nonadjacency.

The same prediction is also counterintuitive vis-à-vis making direct analogies between AGL/ALL paradigms and theories of language acquisition. Empirical data from the acquisition literature suggest that the more frequently children hear a verb used in a given construction, the more firmly its usage becomes *entrenched*, and hence the less likely they will be to generalise that verb to any novel construction with which they have not heard it (Brooks & Tomasello, 1999; Brooks, Tomasello, Lewis & Dodson, 1999). If applied to my AGL paradigm the entrenchment hypothesis predicts that given the large number of repetitions (144) of each of the three sentences in the artificial language, memorising trigrams would seem the most effective way to encode the grammar and generalisation would be hindered. Indeed, Tomasello has argued that an early stage of language acquisition consists of largely unanalysed holophrases, i.e. sentences whose components are not been extracted partly or entirely. The extent to which we can draw parallels between my AGL results and naturalistic or experimental data with children is to be investigated further, but clearly AGL experiments are informative in that they can help promote or demote hypotheses about language acquisition. Below I present an experiment that tests generalisation to new embeddings under conditions of variability. Again, the

experiment below was conducted on adults as a preliminary investigation to be later extended to infants. Subsequently, given the successful results of SRNs in chapter 3 in modeling the variability effect, I present a simulation of the adult data to test the extent to which SRNs can generalise to novel items under conditions of variability.

Experiment 5 (Human data)

Method

Participants

Thirty-six undergraduate and postgraduate students at the University of Warwick participated and were paid £3 each. None of them had participated in previous experiments.

Materials

The training stimuli were identical to those used in Experiment 1, They were presented auditorily via loudspeakers located next to the computer screen. The test stimuli consisted of 12 strings randomised. 6 strings were grammatical and six were ungrammatical. The ungrammatical strings were constructed as in previous experiments by breaking the correct nonadjacent dependencies and associating a head to an incorrectly associated tail. Six strings (three grammatical and three ungrammatical) contained a previously heard embedding, while 6 strings (again

three grammatical and three ungrammatical) contained a new, unheard embedding.

Procedure

Exactly the same procedure as in Experiment 1 was used. Participants sat and listened to the strings. Training lasted approximately 18 minutes, again as in Experiment 1. After training, test instructions were the same except that they contained an additional sentence stating that the strings they were going to hear may contain new words and they should base their judgement on whether the sentence was grammatical or not on the basis of their knowledge of the grammar. This is to guarantee that participant did not select as ungrammatical all the sentences with novel words simply because they contained novel words (Rebecca Gómez, personal communication).

Results and discussion

A polynomial trend analysis showed a significant quadratic effect, $F(1, 35) = 7.407$, $p < .01$. Figure 12 presents the percentage of endorsements for total accuracy in each of the three set-size conditions. No other effect was found. These findings suggest that people generalise at endpoints of the variability spectrum, in the same conditions in which they detect the invariant nonadjacent structure. Thus they support the counterintuitive prediction set out at the beginning of the chapter, namely that detecting invariant frames and generalising to novel slots may be supported by the same mechanisms.

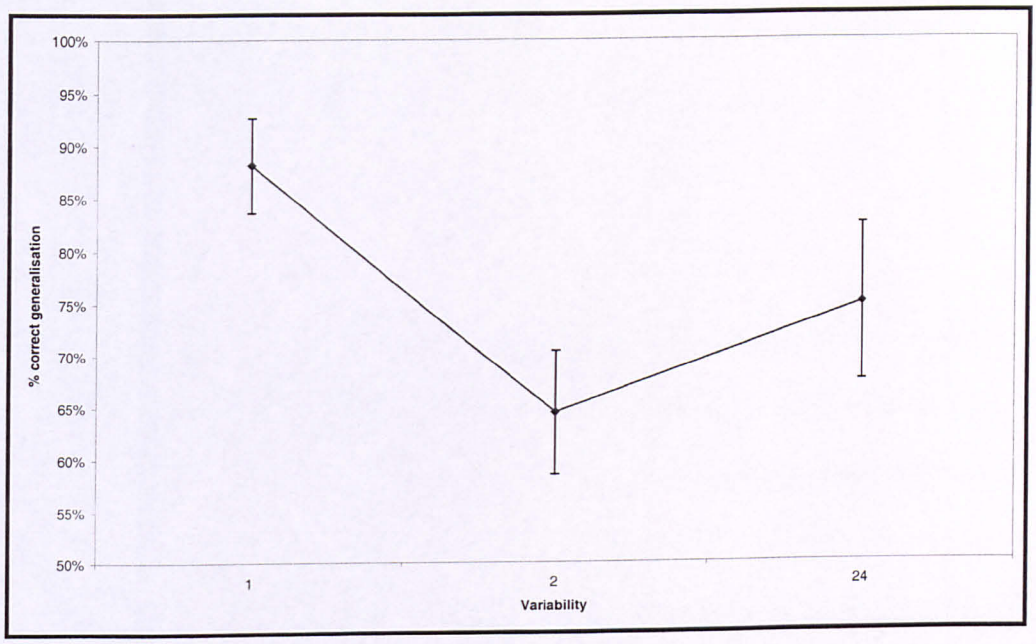

Figure 12. Percent accuracy in generalising to a new embedding across 3 conditions of variability: null, small, and large.

Simulation 3 (SRN data)

The remarkable similarity between the human data in Experiment 1 (chapter 2) and the connectionist simulations in chapter 3 that model a U shape curve in detecting nonadjacencies as a function of variability prompted another series of simulations, which I report below, that attempt to simulate the human data above on generalisation in Simple Recurrent Networks. In principle, if non-local dependencies serve as the backbone for extracting category membership of the embedding, we would expect simple associative mechanisms to master both non-local dependencies and generalisation under the same specific conditions of variability as obtained with the human data. The following simulations test

whether SRNs generalise well in Set Size conditions 1 and 24 but not in Set Size condition 2. The results will be plotted against the human data.

Method

Networks

36 SRNs were trained and tested, corresponding to 12 subjects in each of the 3 conditions in Experiment 5 above. The networks displayed the same structure and initialisation parameters as those in Simulation 2 in chapter 3, except that they contained an extra input and output unit. Again, input and output vector representations were localist, so the new input/output pair served to activate a new middle item to be presented at test. Structurally the networks are equivalent to the networks used earlier.

Materials

Training stimuli consisted of the strings generated from the finite-state grammar used in Simulation 1 and 2 (chapter 2). Test stimuli consisted of 3 previously encountered grammatical strings and 3 ungrammatical strings constructed with previously encountered bigrams (these stimuli were exactly the same as those used in Experiment 2). In addition, another 3 new grammatical and 3 new ungrammatical strings were presented containing previously encountered nonadjacent frames with a new middle item. Such new middle item was represented by activating a localist vector where all the units were off except the new input unit, which had remained off during training. Hence the networks had not encountered this new vector during training.

Procedure

The networks were trained in exactly the same way as in Simulation 2 (chapter 3). The 12 test items were randomised for each network. Performance was measured as in Simulations 1 and 2 in chapter 3, i.e. by calculating the Luce Ratio for the target output node corresponding to the correct tail element of each sentence and turning it into the p-grammatical value.

Results and Analyses

The results are plotted in Figure 13 against human data results. Two out of three expected outcomes were obtained: firstly, generalisation in Set Size 1 was good. Secondly, generalisation in Set Size 2 was worse relatively to Set Size 1, as expected. However, generalisation in Set Size 24 did not improve to levels similar to Set Size 1 and, in any case, was not considerably better than Set Size 2. Hence, the U shape on generalisation can be simulated only partially. This result is puzzling, and is not fully understood at present.

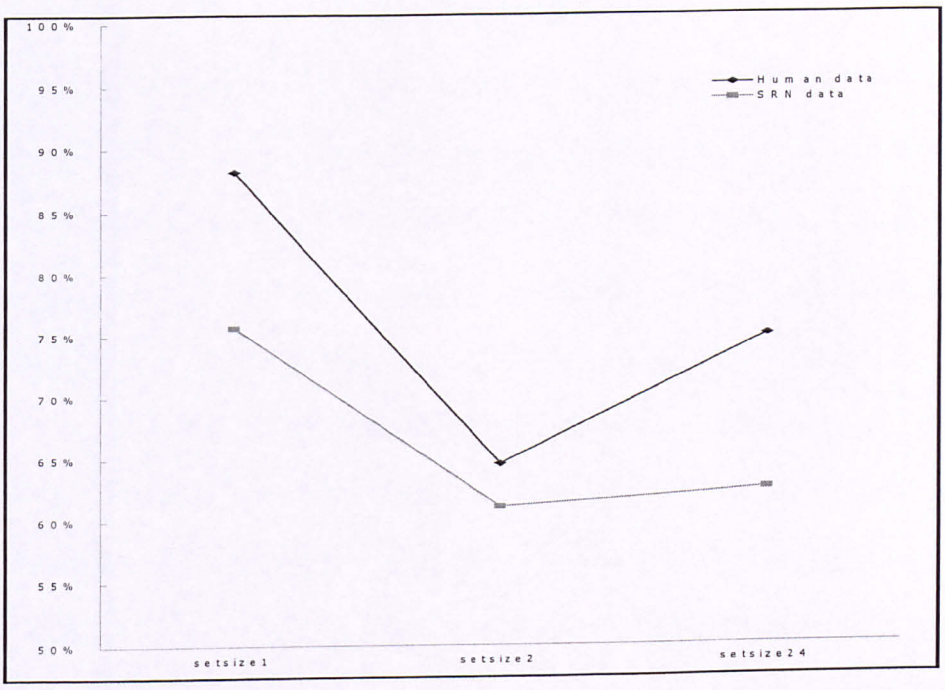

Figure 13. Results from Simulation 3 on generalisation to new embeddings plotted against results obtained experimentally in Experiment 5.

General Discussion

Despite the fact that the potential usefulness of distributional learning has been discounted in the past as irrelevant for learning abstract and syntactic properties of natural languages (e.g. Pinker, 1984), recent times have seen an upsurge of computational and experimental studies investigating the role of distributional learning as part of the larger research endeavour in statistical learning.

In this chapter I have extended the paradigm on detection of invariant structure successfully used in Chapter 2 to investigate whether the process of generalisation of the embedded material is supported under the same conditions of variability. The theoretical issue at stake in this chapter was: do human learners engage distributional learning mechanisms to induce the grammatical category of a word when sufficient and consistent contextual information is given in the input? Extensive statistical analyses of large corpora of child-directed speech strongly support the idea that, in principle, a probabilistic learner would successfully detect the syntactic category structure of words in a language by performing a distributional analysis of the raw input. However, it remained to be ascertained whether learners, adults and children, actually engage in distributional learning. A preliminary step is to test adult learners. Artificial grammars using two-word sentences to elicit word categorisation using lexical co-occurrence patterns have been successful to the extent that a portion of the words are marked by extra cues. However, Mintz (2002) proposed that two-word sentences may not provide sufficient contextual information to learners to engage successful distributional mechanisms because there is no statistical information as to which word constitutes the context and which word constitutes the target to be generalised. Using an *A_X_B* language of the same kind used by Gómez and

here in chapter 2, Mintz found that distributional categorisation was possible when the to-be-generalised embeddings were shared by four contextual frames. By framing generalisation in terms of the variability hypothesis this work has expanded both on chapter 2 and on Mintz's results.

Remember that the learning task in the *A_X_B* grammar was seen in chapter 2 as a question of tracking sequential non-local dependencies in the presence of different degrees of variability of embedded material. Using a figure/ground metaphor, the issue was how to detect non-local invariant structure- metaphorically the figure- upon an invariable or an increasingly variable ground. Conversely, the task tackled in this chapter is how to build a category for embedded material – the figure – and generalise to newly encountered embeddings when such embeddings are shared by several contextual frames, in the same conditions of variability. Just like Rubin's famous face-vase figure (Rubin, 1915), which can be perceived alternatively either as a white vase on a black background or as two black faces looking at each other, in front of a white ground, it is argued that generalising the frame or detecting the embedding are inextricably tied because one leads to the other and vice-versa. The change in perspective, frame versus embedding, I would argue, may lie in the eye of the beholder, in this case the psychologist, rather than strictly being a psychological phenomenon. It is perhaps not psychological in the sense that the same mechanisms, I would argue, are at play in detecting invariant structure versus generalising the embedded material. Indeed, the major contribution of my results is that generalisation to a category EMBEDDING is modulated by the same variability constraints imposed on detecting the frames. Knowledge of this category leads to an abstract representation where a newly heard word can

occupy the embedding slot. Hence, frame detection and generalisation within frames appear to be the two facets of a same distributional process. The crucial inductive problem is, as Mintz (2002) noted, that the learner does not know a priori whether a given word functions as part of a static categorising environment or as a word-to-be-categorised. A completely unprincipled distributional analysis of the input seems cognitively implausible for small artificial grammars, let alone for scaled-up, full-blown language. Perhaps then, learners are naturally biased towards change versus non-change (Gibson, 1991; Gómez, 2002) and this intuition can be formalised in the reduction of uncertainty principle.

In addition to human experiments, this chapter investigated whether basic associative mechanisms as instantiated in Simple Recurrent Networks can replicate the U-shape in the generalisation task just as they replicated so well the U-shape in detecting non-local dependencies in chapter 3. Unfortunately, the picture is not clearcut: although good results under no variability and low scores under small variability were replicated respectively, the average networks' performance with large variability was at 63%, slightly but not tremendously better than the 61% score of networks in low variability. One possible explanation is that the input and output vectors use localist instead of distributed representations. This way of coding the input/output matching may not be conducive to correct classification in neural networks. From the point of view of the network, the new middle item is a completely new vector that bears no resemblance whatsoever with previous vectors. This is equivalent, in the human experimental setting, to showing a completely unrelated item as new embedding at test, say the picture of a cow. It is fairly safe to assume that human participants would have a hard time deciding whether the pseudo-sentence *pel_<picture of a*

cow>_rud was grammatical, regardless of the correctness of the frame. Hence, distributed representations may be a better way of encoding the stimuli in a psychologically plausible way, by representing at least some features of the stimuli common to all other stimuli, for instance phonological properties.

A comment is in order as to the perceptual structure of the stimuli used both here and in chapter 2. To the extent that the middle words contain two syllables versus the one-syllabled heads and tails, the middle words are perceptually augmented by an extra cue. This is because the original Gómez (2002) was devised for children and was meant to maximise perception of the single words in the language (Gómez, personal communication). It could be argued that this study is not dissimilar to the ones mentioned earlier on that utilise distributional information in conjunction with extra cues. As a disclaimer, because the extra-syllable cue is present in all five conditions of variability, I would argue that the crucial factor in both non-local frame detection and embedding generalisation remains uncontroversially the variability effect. In addition, it is well known that natural languages are abundant with phonological, and suprasegmental perceptual cues and a reduction of uncertainty hypothesis does not have to restrict useful information to distributional information. On the contrary, the larger statistical language learning picture one would hope to draw is that learners capitalise on any statistically reliable cues. Recent studies have already established that learners may integrate cues from different domains in their search for structure (Monaghan, Chater, & Christiansen, submitted). In other cases, it may even be the case that some perceptual cues such as stress, preference for certain phonemes or pauses in the speech stream may override

distributional information altogether. This will be the topic of the following chapter.

Lastly, regarding the extension of the present AGL results to the acquisition literature, there is ample scope for debate whether children actually perceive frames and generalise at the same time. Recent work has emphasised the constructive role of syntactic frames as the first step for building more abstract syntactic representations (Gleitman, 1990; Gleitman, Gleitman, Landay & Wanner, 1998; Lieven, Pine, & Baldwin, 1997; Olguin & Tomasello, 1993, Tomasello, 1992, 2000). The most explicitly formulated among these studies (see Tomasello & Brooks, 1999; Tomasello, 2003 for an overview) propose that children's syntactic development builds upon several consecutive stages from holophrases such as *I-wanna-see-it* (at around 12 months), to pivot-schemas (*throw-ball*, *throw-can*, *throw-pillow*, at about 18 months), through item-based constructions (*John hugs Mary*, *Mary hugs John*, at about 24 months), to full abstract syntactic constructions (*a X*, *the Xs*, *Eat a X*). Whether it is possible to closely replicate such developmental patterns using artificial grammars is an open question. One way of doing this is by exposing adult and infant learners to artificial grammars that gradually contain more and more data or that gradually increase complexity and see whether at different stages learners converge towards underlying structures of increasing abstractness and complexity. In general, current artificial grammar experiments with adults and children have so far been limited to the formal aspects of language and these have not been grounded in the functional, pragmatic, and semantic aspects that cover such an important part of language development. For instance, it is implausible to assess whether learners acquire thematic roles such as agent/patient or syntactic

relational categories such as subject/object without matching the meaningless perceptual stimuli to, for instance, objects organised in a visual scene. To my knowledge very few experiments have been conducted in this way (Morten Christiansen has unpublished data, personal communication), so there is ample scope in the future for extending the AGL paradigm to reproduce natural languages more closely.

In the chapters presented so far it has been proposed that statistical learning may be powerful enough to deal with complex sequential stimuli including detecting nonadjacent structure and generalising to novel stimuli. As mentioned at the beginning of this chapter, a main tenet of standard generative linguistic theory and standard cognitive science is that structural linguistic representations are instantiated in the brain as formal algebraic rules. The next chapter will dwell on the nature of generalisations and on the empirical bases to support claims of the distinction between statistical and algebraic computations. It is anticipated that the data presented here and in chapter 2 will be further discussed in chapter 6 in the light of results on speech segmentation using nonadjacent structures. It will be possible to directly compare the two experimental paradigms because the structure of the artificial grammars used is very similar, namely it involves the now famous 3 nonadjacent dependencies with a number of intervening items.

Chapter 6

The debate over the nature of linguistic representations

What computational processes are implicated in language acquisition, and how might we assess them? One recent debate has centered on the extent to which language acquisition is dependent on the statistical structure of the language environment, or on algebraic, rule-like computations (Marcus, 1999; McClelland & Plaut, 1999; Hahn & Chater, 1998). This question has been central to debates about language acquisition, and is ubiquitous at all levels of description of language structure.

The traditional view of language acquisition holds that statistical computations may be useful for learning the sounds and the lexicon of a specific language, but that they are not central to the characterization of grammar, i.e., the set of abstract and universal properties of the language faculty (Chomsky, 1957; Pinker, 1989). At the level of speech segmentation, statistical distributional information might provide information about word boundaries. For instance, in the second half of their first year children begin to distinguish strings of sounds containing legal sequences of sounds - phonotactic constraints - in their language from illegal sequences (/zw/ and /vl/ appear at the beginning of words in Dutch but not in English; Jusczyk, 1999). Infants are also capable of exploiting statistical regularities as cues to speech segmentation. For instance, when hearing a continuous stream of syllables both adults and children were sensitive to points where transitional probabilities between speech sounds were lowest (Aslin, Saffran, & Newport, 1996; Saffran, Aslin, & Newport, 1996). At the word-level, connectionist models, which pick up on distributional information in the environment, indicate that statistical information may play a large role in determining mappings between written and spoken forms of words and their

meaning (Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989). Similarly, at the grammatical level, connectionist models have renewed interest in the language structure that can be learned from distributional statistics (Christiansen & Chater, 1999; Elman, 1990). The shifting balance between statistical and rule-like approaches in language modeling can also be observed in the changing emphasis between symbolic and statistical methods in computational linguistics (Klavans & Resnik, 1996; Manning & Shütze, 2000). In this area of studies successful integrative approaches to language have been used rather than a commitment to either purely symbolic or statistical methods. The core research topic is the probabilistic nature of language at all levels of analysis (comprehension, production, phonology, morphology, syntax, semantics, sociolinguistics; see Bod, Hay, & Jannedy, 2003).

Peña, Bonatti, Nespor, and Mehler (2002) provided a set of experiments that, they argued, showed that such views could be reconciled: speech segmentation operates on the basis of statistical learning, whereas entirely separate algebraic computations are necessary for learning grammatical structure. In this chapter I present a series of experiments to show that this line of evidence does not yet support this segregation of computational processes. I discuss methodological questions pertaining to the merits and limits of Artificial Language Learning (henceforth ALL) experiments as tools of investigation, and caution against theoretical conclusions based on tests without full controls. In addition, I evaluate the limits of casting the debate on language learning using the current rules versus statistics dichotomy. Recent ALL studies point both to a general natural predisposition to discover structure, in whatever form, and to a richness of potential cues that both children and adults can exploit. Key issues

involve discovering the principles that guide learners to choose among competing sources of information, to integrate or discard them. My results suggest that salience and learnability of a particular structure may be heavily dependent upon the perceptual and probabilistic factors as well as the training and test conditions.

Are algebraic and statistical computations empirically separable?

Since the first studies using artificial grammars (e.g. Miller, 1967; Reber, 1967), convincing evidence has been accumulated that adults become sensitive to the deep structure contained in chained events such as strings of letters, sounds, or images. This line of research has been successfully extended to infants (see for instance Gómez, 1999; Jusczyk, 1999; Saffran, Aslin, & Newport, 1996). This learning usually takes place after limited and incidental exposure to complex stimuli. In a typical ALL situation, participants are first exposed to numerous stimuli and asked to memorize or process them in some way. Subsequently, they are informed that the stimuli were generated by a specific set of rules (a grammar), and are asked to classify further strings as grammatical or not. Typically, participants achieve some degree of success in this classification task despite their limited ability to recognize or verbalize overtly the knowledge of the features that define grammaticality. The learning mechanisms involved in such situations remain controversial. Nonetheless, ALL paradigms have been used as an empirical test-bed of the statistical versus algebraic debate in language acquisition. Knowledge of algebraic rules is characterized as the representation of mental abstract variables. Symbolic accounts of language (Marcus, 2001; Pinker, 1999) define linguistic processes as operating over such variables.

It has been claimed that rule knowledge is necessary in order to afford limitless linguistic generalizations across the board to any novel item, regardless of familiarity with the features of previously encountered items (Marcus, 2001). For instance, speakers are able to generalise regular inflection such as the *–ed* suffix marking the past tense in English to novel and strange-sounding words (Prasada & Pinker, 1993). Several connectionist models of inflection eliminate the representation for variables like "stem" and operations like "stem+s" in the formation of English plurals (e.g. Daugherty & Seidenberg, 1992; Hahn & Nakisa, 2000; Rumelhart & McClelland, 1986; Plunkett & Juola, 1999). However, Marcus argued that the kind of generalization subserved by connectionist models is limited compared to human generalisation: although connectionist networks can generalise to a novel item that bears resemblance to the trained regular instances (e.g. they can generalise to a novel noun *blick* because of familiarity with previously encountered nouns *brick* or *block*), unlike humans they fail to generalise to novel nouns like *xick* (pronounced /xIk/) whose features fall outside the training space, as defined by Marcus (Marcus, 1998, 2001).

Conversely, statistical language learning studies have highlighted human learners' sensitivity to statistical properties of the input. Saffran, Aslin, and Newport (1996) familiarised 8-month-olds to a stream of concatenated 3-syllable word-like stimuli, such that the transitional probability of a syllable given the preceding syllable within a word was 1, whereas syllable transitional probabilities crossing word boundaries were .33. At test, they found that infants preferred isolated words as opposed to part-words containing syllables that spanned word boundaries, i.e., there was a preference for stimuli that maximized

the transitional probabilities between syllables. Supporters of statistical language learning argue that the brain is endowed with powerful general statistical computations similar in style to those implemented in connectionist networks. To test this position, ALL experiments have been used to assess participants' ability to learn abstract grammatical structure. Such studies have typically focused on cases where learning occurs where no apparent statistical distributional information is available in the stimuli. In such cases participants are required to abstract the underlying rule from a set of training stimuli to a novel stimulus which obeys the rule of the training set but which has not been seen previously. Such generalisations, however, have been characterized either in terms of rule-learning (Marcus *et al.*, 1999), statistical learning (Gómez & Gerken, 1999), or both (Redington & Chater, 2002). There remains, however, the possibility that both statistical and algebraic computations play a role in language learning.

Peña *et al.* (2002) provided a set of intriguing ALL studies that seemed to suggest that statistical computations are used for segmentation, but cannot be used for learning rules in the language. Rather, rule learning is subserved by a distinct type of computation. Their participants were presented with continuous streams of syllables comprised of words of the form $A_iX_jB_i$, where there were three such $A_i\_B_i$ pairs, and X_j was one of three syllables that randomly intervened between the $A_i\_B_i$ pair. In a subsequent test phase, participants demonstrated a preference for words (e.g., $A_1X_2B_1$) over part-words, i.e., sequences that crossed word boundaries (e.g., $X_2B_1A_3$ or $B_3A_1X_2$). The nonadjacent dependencies between the A_i and the B_i syllables were learned and contributed towards segmentation. Following an identical training phase, Peña *et al.* (2002) tested participants on whether they learned to generalize from the rules of the stimuli.

Participants demonstrated no preference for "rule-words", composed of an $A_i\_B_i$ pair with a different A or B in the intervening position (e.g., $A_1B_3B_1$), compared to part-words.

In a third manipulation, 25-ms gaps were introduced between words during the training phase of the experiment, and now participants generalized as indicated by a preference for rule-words over part-words. Peña *et al.* claimed that altering the speech signal resulted in a change in the computations performed by their participants. Statistical computations were used in a segmentation task but this was not performed simultaneously with algebraic computations that would permit generalizations of the structure. Once the segmentation task was solved by introducing small gaps in the speech signal, the underlying structure would be learned.

An alternative explanation to account for the results is that, as Seidenberg, MacDonald, and Saffran (2002) point out, certain phonological properties of the stimuli may have contributed to preferences for certain words. In each experiment, Peña *et al.* (2002) used syllables in the same positions. In addition, all initial and final syllables began with a stop consonant. It is possible, then, that phonological properties exert an influence on the results rather than that participants learn the subtle statistical or algebraic properties of the stimuli. As a first step I carried out a corpus analysis to investigate the distribution of the consonants used in Peña *et al.*'s experiments. The experiments in Peña *et al.* were performed on French speakers, the experiments I present in this chapter were on English speakers, so I here consider both languages.

I assessed the percentage of words (taking into account their frequency) in the Brulex corpus of French (Content, Mousty, & Radeau, 1990) and in the

CELEX corpus of English (Baayen, Piepenbrock, & Gulikers, 1995) that began with each phoneme from the syllables used by Peña *et al.* The results are shown in Table 5. In French, initial phonemes from Peña *et al.*'s materials were more likely than medial phonemes to begin words, and in English, initial phonemes were more likely than both medial and final consonants to begin words. I tested the consequence of forming a preference for words over part-words based only on the likelihood of the initial phoneme in word-initial position. From the 36 tests of word/part-word in the segmentation experiment in Peña *et al.*'s study, in French 23 cases produced a preference for a word over a part-word, and in English 32 words would be preferred over part-words. In each language, response selection on the basis of onset probability of the latent language would result in highly significant mean responses.

Position	Phoneme	Percentage of onsets in French	Percentage of onsets in English
Initial	/p/	6.67	3.11
	/b/	2.20	4.45
	/t/	5.00	4.89
		total: 13.87	total: 12.45
Medial	French /R/, English /♦/	3.95	2.16
	/f/	4.60	4.36
	/l/	2.02	2.26
		total: 10.57	total: 8.78
Final	/k/	8.92	3.69
	/g/	1.22	1.50
	/d/	6.51	2.99
		total: 16.65	total: 8.18

Table 5. Percentage of words beginning with each consonant for syllables in initial/medial/final word position in Peña *et al.*'s studies.

Seidenberg, MacDonald and Saffran (2002) indicated that, in Peña *et al.*'s stimuli, all initial and final syllables began with a stop consonant, whereas medial syllables began with continuants. Taken together with my corpora analyses, there is mounting evidence that phonology may potentially influence task performance in ALL experiments. Given the effectiveness of responding according to positional frequencies of phonemes from the latent language I

performed a series of experiments to test empirically the extent to which phonology may influence task performance in ALL.

I present below a battery of new ALL experiments that manipulate the order and position of syllables, which indicate that the confound of phonology is sufficient to account for all of the results obtained by Peña *et al.* Consequently, there is no evidence yet for learning, either statistical or algebraic, on the basis of the nonadjacent dependencies in the stimuli. I divide the experiments into sets relating to the segmentation task, as proscribed by Peña *et al.*, and to generalization of structure. Experiments 1 to 3 concern segmentation, and Experiments 4 to 8 explore the issue of generalization. The first three experiments test the extent to which phonology can account for word over part-word preferences. The first experiment precisely replicates Peña *et al.*'s experiment where words were preferred over part-words. The second experiment tests whether the preference for words was due to the particular choice of phonemes in different positions within the words, and the third experiment tests whether the preference for words over part-words pertains when phonemes maintain their position, but the structure is removed.

Experiment	Peña *et al.* experiment	Segmentation/Generalisation task	Syllable positions	*Nonadjacent* Structure	25ms Gap	Effect
6	1	Segmentation	Original	Y	N	< .00001
7		Segmentation	Randomised	Y	N	*ns*
8		Segmentation	Original	Random	N	< .005
9	2	Generalisation	Original	Y	N	ns
10		Generalisation	Randomised	Y	N	< .05*
11	3	Generalisation	Original	Y	Y	< .01
12		Generalisation	Randomised	Y	Y	< .005
13		Generalisation	Original	Random	Y	< .01

Table 6. Summary of the design of the experiments. The first column lists the Experiment, the second column lists the experiment number in Peña *et al.*'s study. "Syllable positions" indicates whether syllables occurred in the original initial/medial/final positions from Peña *et al.* The "Structure" column indicates whether the language contained nonadjacent dependencies or not, and the effect indicates the statistical result (* indicates that there was a significant reverse effect, i.e., there was a preference for part-words over rule-words in Experiment 10).

Experiment 6

In Experiment 6, I wanted to replicate Peña *et al.*'s finding that participants have a preference for words over part-words. I precisely replicated Peña *et al.*'s first experiment except using English participants and utilizing synthesized spoken English.

Method

Participants

10 undergraduate and postgraduate students at the University of Warwick participated for £1. All participants spoke English as a first language and had normal hearing.

Materials and design

We used the same nine word types from Peña *et al.* to construct the training speech stream in Experiment 6. The set of nine words was composed of three groups ($A_i\_B_i$), where the first and the third syllable were paired, with an intervening syllable (X) selected from one of three syllables. The first set (A_1XB_1) was: [pu-li-ki], [pu-ra-ki], [pu-fo-ki]; the second set (A_2XB_2) was: [be-li-ga], [be-ra-ga], [be-fo-ga]; and the third set (A_3XB_3) was: [ta-li-du], [ta-ra-du], [ta-fo-du].

Words were produced in a seamless speech stream, with no two words from the same set occurring adjacently, and no same middle item occurring in adjacent words. I used the Festival speech synthesizer (Black, Taylor, & Caley, 1990) using a voice based on British-English diphones at a pitch of 120 Hz, to generate a continuous speech stream lasting approximately 10 minutes. All

syllables were of equal duration, and were produced at a rate of 4.5 syllables/second. Words were selected randomly, except that no $A_i\_B_i$ pair occurred twice in succession. The speech stream was constructed from 900 words, in which each word occurred approximately 100 times. The speech stream faded in for the first 5 seconds, and faded out for the last 5 seconds, so there was no abrupt start or end to the stream.

Part-words were formed from the last syllable of one word and two syllables from the following word (B_iA_jX), or from the last two syllables of one word and the first syllable from the following word (XB_iA_j). Participants were seated in individual sound-proof labs. E-prime was used to present training and test speech, which was played through centrally-positioned loudspeakers.

Procedure

In the training phase, participants were instructed to listen to continuous speech and try and work out the "words" that it contains. They then listened to the training speech. In the testing phase, participants were requested to respond which of two sounds was a "word" in the language they had listened to. They were then played a word and a part-word separated by 500 ms, and responded by pressing either "1" on a computer keyboard for the first sound a word, or "2" for the second sound a word. After 2 seconds, the next word and part-word pair were played. In half of the test trials, the word occurred first. 5 participants heard a set of test trials with one set of words first, and the other 5 participants heard the other set of words first.

Results

The results are reported in Figure 14. The top part of the Figure represents a sample of the training phase. Colours indicate different words. The rest of the figure reports individual scores (single dots) in preferring words (on the left hand) versus part-words (underscored on the right hand), expressed in averaged percentages. The results replicated those of Peña *et al.* Participants preferred words over part-words, with a mean score of 29.3 (81%) from a possible 36, where chance performance equals 18. A single-sample *t* test (two-tailed) showed overall performance significantly better for words over part-words: $t(9) = 6.81$, $p < .001$. In addition, participants preferred words significantly more when they had to make a decision against part-words of the form XB_iA_j (the mean score was 15.9 from a possible 18) as opposed to part-words of the form B_iA_jX (the mean score was 13.4 from 18), $t(9) = -2.82$, $p < .05$.

Discussion

The replication of Peña *et al.* is a preliminary requisite to ensure direct comparison between the task being carried out on English and French participants. I found that, even though the language and the synthesizer differed from that of the experiments on French, the same strong preferences for words over part-words were found in my study. Given the similarity between the distribution of plosives in English and French – plosives occur word-initially more than laterals – there remains the possibility that participants are guided in their responses by the distributions in their latent language rather than by the structure of the artificial language. Additional evidence for the possibility of phoneme preferences influencing the results comes from the significant

differences in preferences for words over B_iA_jX part-words compared to words over XB_iA_j part-words. B_iA_jX part-words began with a plosive, and thus exhibited a small preference for words. In contrast, XB_iA_j part-words began with a lateral or a fricative, and words, beginning with a plosive, were much preferred over these part-words.

There is thus a distinct possibility that the word over part-word preferences exhibited in Experiment 6 were due to preferences for phonemes in certain positions. In order to test this possibility, I ran a control version of this study that broke the link between certain phonemes occurring in initial, medial, or final positions in Experiment 7. An additional source of preference for words over part-words was that words occur approximately twice as frequently in the training speech corpus as part-words. I control for this potential influence on the results in Experiment 7.

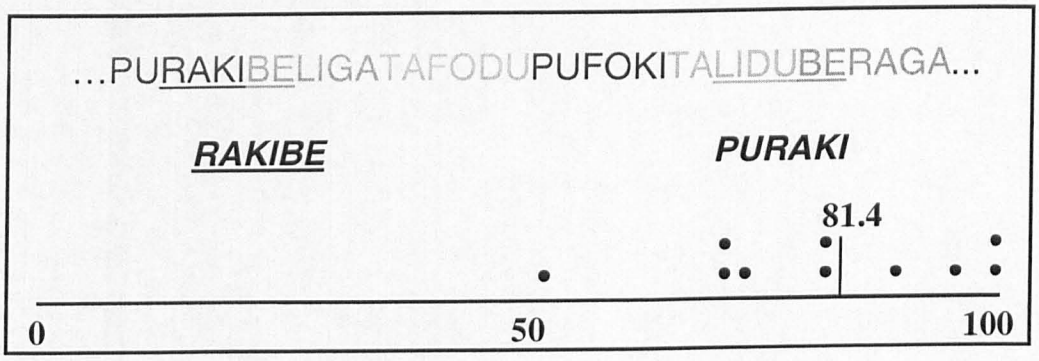

Figure 14. At the top of the frame, a sample of the training speech is shown, with "words" shown in different colours and part-words underlined. Underneath, is a sample of a test pair: in Experiments 6-8, words were compared to part-words, in 9-13, rule-words were compared to part-words. The results for each participant, in terms of percentage preference for part-word or word/rule-word, is represented by a dot. The mean for all participants is indicated above a vertical line. Experiment 6 – segmentation task.

Experiment 7

In Experiment 7 I tested whether performance was guided by preference for syllables beginning with /p/, /b/, or /t/ in word-initial position. In order to test for this preference, for each participant I randomly assigned each of the nine syllables in Experiment 6 to three $A_i\_B_i$ pairs and three X's. Each participant was exposed to a training corpus that had the same structure, but with phonemes assigned to different positions.

Method

Participants

10 students from the same population, but who had not participated in any other experiment reported here, participated for a £1 payment.

Materials and design

For each participant, I randomly assigned 6 of the syllables from the first experiment to the $A_i\_B_i$ pairs, and the other three syllables to the X_j position. Thus, each participant listened to speech with the same structure containing the nonadjacent dependencies, but with syllables assigned to different positions. For instance, the sequence $A_1X_3B_1$ was instantiated as [li-ki-pu] for one participant but as [ra-be-ga] for another one. Once the syllables had been assigned to the positions within the words they remained in those positions for the duration of the experiment. In addition, because part-words were half as frequent as words in the training phase in Experiment 6, I doubled the frequency of one of the three $A_i\_B_i$ pairs and then used the other two words compared to part-words comprised

of the first or last phoneme of the higher-frequency word together with two syllables from a lower-frequency word. In this way, both word and part-word sequences at test had been heard with the same frequency. Test items were composed of one of the lower-frequency A_iXB_i words and either a XB_iA_j or a B_iA_jX part-word, where B_j and A_j were from the higher-frequency word. All 12 possible word and part-word pairs were used, and participants responded to 24 pairs, 12 of which had the word preceding the part-word, and 12 in which the part-word preceded the word.

Procedure

The training and testing procedure were identical to that for Experiment 6.

Results

The results are shown in Figure 15. No preference was found for words over part-words. The mean response correct was 11.4 (47%) from a total of 24, which was not significantly different from chance, $t(9) = -0.56$, $p = .58$.

Discussion

The results for Experiment 7 contrast with those of Experiment 6 strikingly. The key change that I made between Experiment 6 and Experiment 7 was to reassign syllables to different roles for each participant. The structure of the language was identical for both Experiment 6 and Experiment 7, however the strong preferences for words over part-words observed in Experiment 6 were completely absent from Experiment 7. That is, when preference for phonemes in

onset positions was controlled there was no indication of learning the nonadjacent dependencies in the speech signal[10]. This provides strong evidence for rejecting the hypothesis that participants learn the underlying structure of the language and use this to guide their preferences for certain sequences of speech sounds.

This lead us to run a further control in Experiment 8, where I remove the nonadjacent dependency structure from the language but maintain the original phonological positions of syllables. This tests whether phonological preference alone is sufficient to determine preference for one guides performance.

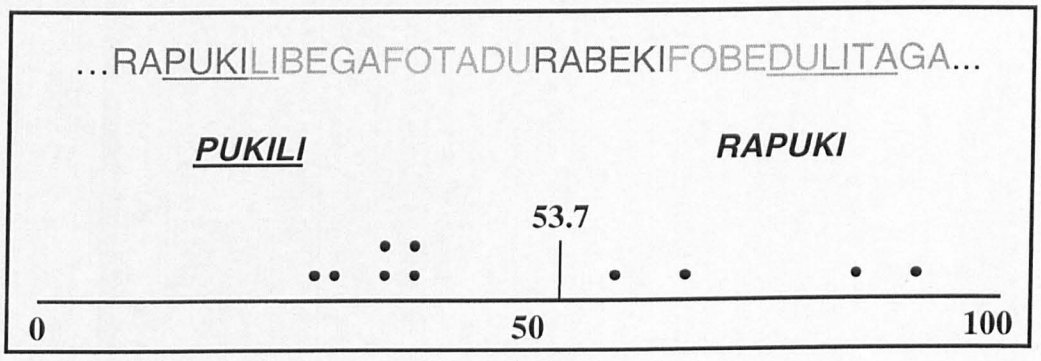

Figure 15. Experiment 7 – segmentation task with randomized phonology.

[10] Peña *et al.* repeated their experiment 1 by interchanging part-words for words during the training phase. They found a reduced, but significant, preference for words over part-words. We suggest that testing a single control is not sufficient for removing any phoneme positional preferences (note that in 4/10 cases in our Experiment 7, participants performed at better than chance level which may have reflected respecting the phonological preferences of the latent language in 4/10 randomizations of the syllable ordering).

Experiment 8

If preference for phonemes in certain positions is the explanation for decisions at test, then we should also find a significant preference for words over part-words when the phoneme positions are as in Peña *et al.* (2002), but nonadjacent structure is removed. Experiment 8 tested whether participants would prefer phonemes in particular orders even when there was no nonadjacent structure in the speech signal. I maintained the order of phonemes from Experiment 6, but broke the dependency between the first and the third syllable in each word. So, any first syllable was followed by any second syllable, which could be followed by any third syllable.

Method

Participants. 10 students (who had not participated in any other experiment reported here) at the University of Warwick participated for £1.

Materials and design.

The methods in Experiment 8 were the same as for Experiment 6. The speech stream differed in that the 9 syllables of Experiment 6 maintained their relative positions within words, but any combination of A, X, and B could occur within a word. For instance, whereas in Experiment 6 the first syllable [pu] was always paired with the last syllable [ki], generating a nonadjacent frame [pu-X-ki], now it generated two more frames [pu-X-ga], and [pu-X-du]; likewise for the other syllables. Hence, the speech stream was now comprised of 27 word types, and each word occurred approximately 33 times in the speech stream in randomized order with the constrain that no adjacent two words shared first, second, or third

syllable. The test phase consisted of all 27 words, compared to part-words that were composed of either the last two syllables of the word followed by the first syllable of the word (e.g., the word A_iXB_j was compared to the part-word B_jA_iX or XB_jA_i).

Procedure

The training and testing procedure were identical to that for Experiment 6 in every other way.

Results

The results are shown in Figure 16. Participants in this Experiment preferred words over part-words with a mean of 17.2 (63%) from a total of 27, which was significantly different from chance, $t(9) = 4.20$, $p < .005$. There was no difference in responses to B_jA_iX or XB_jA_i part-words.

Discussion

The results of Experiment 8 indicate that, even though there was no structure at all in the artificial language, participants still exhibited a preference for words over part-words, as defined by positions of phonemes. Taken together, Experiments 1 to 3 provide strong evidence that participants have not learned to solve the task based on learning nonadjacent dependencies. Experiment 7, which maintained the nonadjacent structure from Experiment 6, but randomized assignment of syllables to particular positions for each participant, found no evidence for learning. Experiment 8, which had no structure, but maintained

syllable positions from Experiment 6, found a significant preference for words over part-words. Words, in this case, are not defined by the structure of the artificial language, but rather by phonological information. I conclude that there is, as yet, no empirical evidence in support of Peña *et al.*'s claims that nonadjacent dependencies are helpful for segmentation.

I have as yet found no evidence for the learning of nonadjacent dependencies, but I have found profound influences of phonological preferences on task performance. I next assessed the extent to which studies purporting to show learning generalizations can be accounted for in terms of phonological preferences rather than the learning of nonadjacent structure.

Peña *et al.* (2002) tested whether participants identified the structural nonadjacent dependencies when presented with novel strings that contained the previously seen dependencies and a new intervening middle item. To do this, they tested participants' preference for part-words versus "rule-words". These were defined as words whose medial syllable was taken from another $A_i\_B_i$ pair. For instance, having heard [pu-li-ki], [pu-ra-ki], and [pu-fo-ki] during training participants where tested on a new sequence [pu-be-ki], with the syllable *be* having occurred as either the initial or final element of another $A_i\_B_i$ pair. Experiment 9 tests participants' preferences for rule-words over part-words when the training speech corpus was identical to that of Experiment 6. Experiment 10 tests whether the same results emerge when syllables are randomly assigned to different positions for each participant. Experiment 11 tests the influence of introducing a short gap in the speech between words in the training corpus. This was found to produce a preference for rule-words over part-words in Peña *et al.*'s studies when this was not found without a gap. Experiments 7 and 8 test this

effect when syllables are randomly assigned to different positions, and when the structure is removed but syllables maintain their original positions within words.

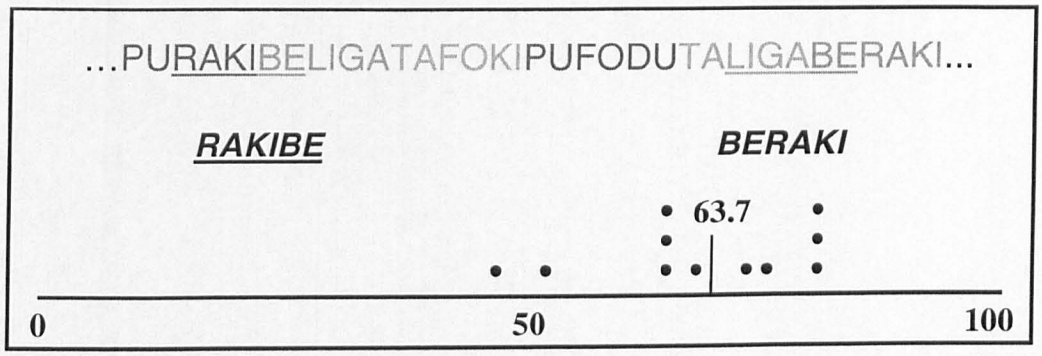

Figure 16. Experiment 8 – segmentation task with no structure.

Experiment 9

We used precisely the same training stimuli as in Experiment 6, but tested participants' preference for "rule-words" compared to part-words. Peña *et al.* predicted that distributional information alone could not afford this generalization and hence rule-words should not be preferred to part-words. Experiment 9 was a replication of Peña *et al.*'s Experiment 7, but with English participants and English synthesized speech.

Method

Participants

10 students (who had not participated in any other experiment reported here) at the University of Warwick participated for £1.

Materials and design

Experiment 9 was identical to Experiment 6 except for the test items in the test phase. Part-words were now compared to "rule-words", which were composed of $A_i\_B_i$ pairs with an intervening item that was either an A_j or a B_j from another $A_j\_B_j$ pair. I used the same rule-words as Peña *et al.*: for the A_1XB_1 set the rule-words were [pu-be-ki], [pu-ta-ki], [pu-ga-ki]; rule-words for the A_2XB_2 set were [be-du-ga], [be-ki-ga], [be-pu-ga]; and [ta-ga-du], [ta-be-du], [ta-ki-du] for the A_3XB_3 set. Part-words were constructed in the same way as in Experiment 6, and there were 36 test items.

Procedure

The training and testing procedure was identical to that for Experiment 6 in every other way.

Results

The results are shown in Figure 17. In line with Peña *et al.*, I found no evidence for participants learning to generalize from the nonadjacent structure of the stimuli. Participants responded with a preference for rule-words over part-words 17.1 (47%) times from a total of 36. This was not significantly different to chance, $t(9) = -.55$, $p = .59$.

Discussion

We found no evidence for a preference for rule-words to part-words, which replicates the results of Peña *et al.* precisely. These negative results were interpreted by Peña *et al.* as decisive evidence that "a computational mechanism sufficiently powerful to support segmentation on the basis of nonadjacent transitional probabilities is insufficient to support the discovery of the underlying grammatical-like regularity embedded in a continuous speech stream"(p.606). It is possible that the lack of preference for rule-words over part-words, or vice versa, was obscured by the phonological preferences found in Experiments 1 and 3. Experiment 10 tests whether there are preferences when syllables do not occur in particular positions.

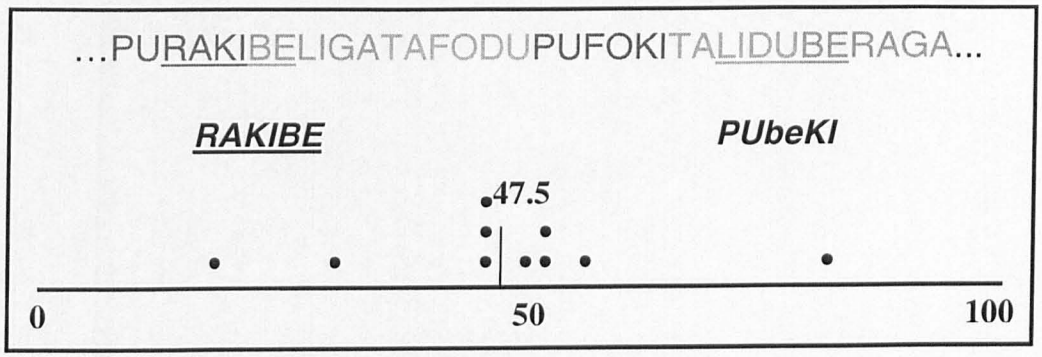

Figure 17. Experiment 9 – generalization task.

Experiment 10

To test the possibility that generalizations were learned, but overridden by preferences for certain orders of phonemes, I randomized the phonology for each of the 10 participants in Experiment 10. The language for each participant was thus generated by assigning syllables to different positions for each participant, but maintaining the nonadjacent structure. The training speech corpora were thus produced in exactly the same way as for Experiment 7 above, except that all words occurred with equal frequency. Then, I tested participants on preference for "rule-words" versus part-words.

Method

Participants

10 students (who had not participated in any other experiment reported here) at the University of Warwick participated for £1.

Materials and design

Experiment 10 was identical to Experiment 9, except that the assignment of syllables to words was different for each participant, similar to the assignment reported in Experiment 7. Rule-words were constructed by taking a syllable from one of the other $A_i\_B_i$ pairs in precisely the same way as for Experiment 9, such that at least one rule-word was composed with an intervening A_i and at least one with a B_i from another nonadjacent pairing. There were 36 test pairs.

Procedure

The training and testing procedure was identical to that for Experiment 9 in every other way.

Results

Surprisingly, I found a preference for part-words over rule-words, as shown in Figure 18. Participants preferred rule-words to part-words a mean 15.1 (41%) times from 36, which was significantly less than chance, $t(9) = -2.73, p < .05$.

Discussion

This control experiment shows that participants preferred sets of syllables that they had heard during the training phase over the "rule-words", which were novel sequences. This preference was overshadowed in Experiment 9 by the preference for certain onset phonemes, and was similar to the preference for familiar sequences found when Peña *et al.* (2002) familiarized their participants to an extended 30 minutes of continuous stream in their Experiment 9. It may be that familiarity with part-words obscures any learning of structure that admits generalization to rule-words. Rule-words are unfamiliar sequences, and the interposition of an element that has been learned to occur in a different position may interfere with learning. However, it remains the case that no evidence for generalization to rule-words was found in Experiment 9 or 5. Experiment 11 tests whether the introduction of a short gap between words changes the computations involved in learning the structure of the language.

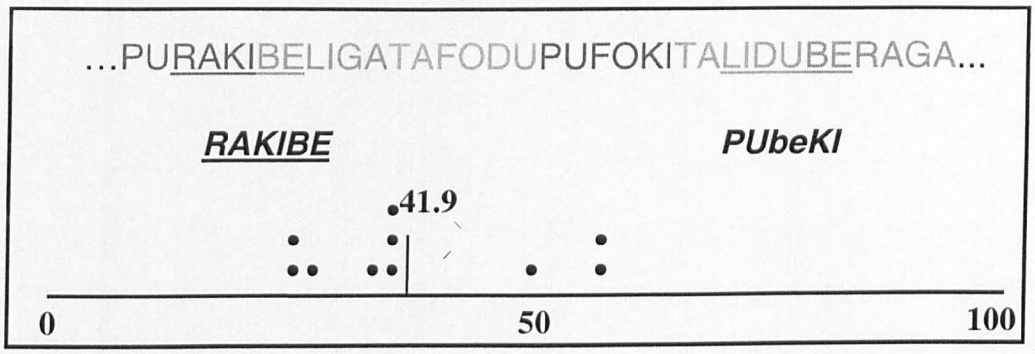

Figure 18. Experiment 10 – generalization task with randomized phonology.

Experiment 11

The next set of experiments (11-13) test whether generalizations occurred when subliminal gaps were introduced between words. In Peña *et al.*'s view, generalisation is not triggered by distributional analysis of the input, but by a different type of signal. The introduction of a subliminal gap was interpreted as relieving participants from the burden of computing transitional probabilities, thus allowing them to capture the generalizations in the language. Experiment 11 replicated Peña *et al.*'s third experiment, which was precisely the same as my Experiment 9 except that gaps of 25 ms intervened between words during the training phase. The Experiment tested whether rule-words would be preferred over part-words when a gap intervened between words in the training speech corpus.

Method

Participants

10 students (who had not participated in any other experiment reported here) at the University of Warwick participated for £1.

Materials and design

Experiment 11 was identical to Experiment 9 except for the training speech stream. Words were now separated by a 25ms pause. The Experiment precisely replicated the third experiment of Peña *et al.* (2002), except that participants were English speakers and the speech synthesizer was based on British English diphones.

Procedure

The training and testing procedure was identical to that for Experiment 9.

Results and discussion

Participants reliably preferred rule-words to part-words, with a mean of 22.8 (63%) preferences for rule-words from 36 items, $t(9) = -3.41$, $p < .01$. This result is consistent with Peña *et al.* and has been taken to suggest that, once the segmentation task has been solved by the introduction of gaps between words, participants are free to concentrate on the structure of the language. Generalizations from this structure, reflected by preferences for rule-words over part-words, were taken to indicate learning the abstract rules of the language. The results of the previous Experiments have cautioned against hasty conclusions based on results from studies that have not controlled for potential phonological preferences. Experiment 12 tested the extent to which preference for certain phonemes in different positions within the word might account for the rule-word preference.

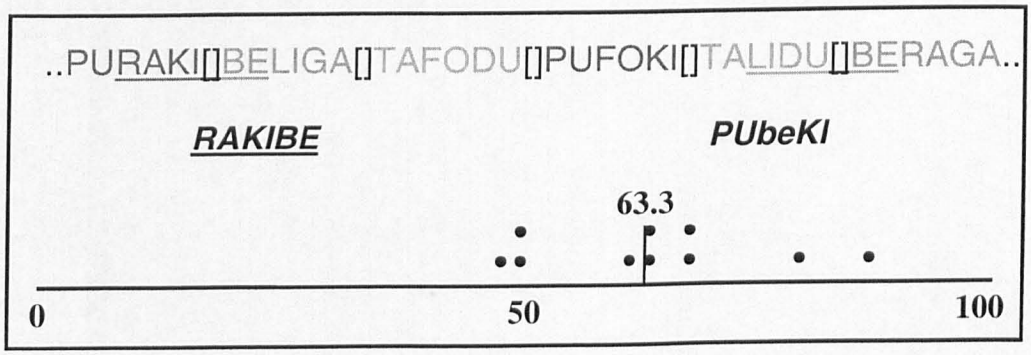

Figure 19. Experiment 11 – generalization task with gap.

Experiment 12

Peña *et al.* argued that the introduction of the gap induced a different type of computation on the speech signal. Experiment 12 tested whether this effect was due to the particular choice of phonemes in Experiment 11 by randomly assigning syllables to the different positions and words in the artificial language for each participant, but maintaining the $A_i\_B_i$ structure of the language.

Method

Participants 10 students (who had not participated in any other experiment reported here) at the University of Warwick participated for £1.

Materials and design

The same randomization used in Experiments 2 and 5 was adopted, such that the structure of the language was maintained, but syllables were randomly assigned to different positions within words for each participant. Training and test stimuli were the same as Experiment 10, except for the introduction of the 25ms gap between words in the training speech corpus.

Procedure

The training and testing procedure was identical to that for Experiment 11 in every other way.

Results

The results are shown in Figure 21. Participants responded with a preference for rule-words over part-words with a mean of 24.9 (69%) from 36 responses, which was significantly greater than chance, $t(9) = -4.40$, $p < .005$.

Discussion

Experiment 12 tested whether there was a preference for rule-words over part-words when preference for phonemes in particular positions was controlled for. I found that this was the case – rule words were preferred to part-words to a significant extent. There are two possible explanations for this effect. First, it may be that, as Peña *et al.* claim, generalizations are learned by participants. In this case, phonological preferences cannot account for the results. The second explanation, as noted by Seidenberg *et al.*, is that the gap adds salience to the initial syllables, meaning that preference for previously heard words is over-ruled by the novel words beginning with the salient initial syllables. To test the subliminal status of the gaps, Peña *et al.* ran a control (note 22, Peña *et al*, 2002) where they played two sequences of 1 minute from the artificial language, one of which contained the gaps. Afterwards, participants were informed of the gap, and asked whether they had noticed any difference in the two sequences they had just heard. They were asked which of the two sequences contained them, and responses were at chance.

According to Holender (1986) a more conservative way to test for subliminality is to inform participants of the subliminal element *before* presenting the stimuli. I performed this more conservative test using the same sound files that Peña *et al.* used. 10/10 participants identified the sequence with

gaps correctly. Under this stricter test, the gaps proposed by Peña *et al.* can no longer be said to be subliminal. The possibility therefore remains that learners were not detecting the nonadjacent structure, but rather were responding according to the salience of the first syllable, induced by the introduction of a gap prior to the syllable. If this is the case, then participants should still indicate a preference for rule-words versus part-words even in the absence of nonadjacent structure. I tested this possibility in Experiment 13.

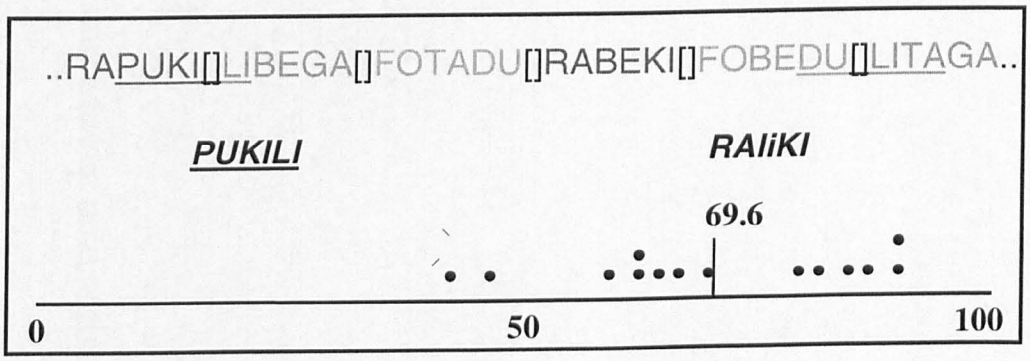

Figure 20. Experiment 12 – generalization task with gap and randomized phonology.

Experiment 13

We tested the extent to which performance was due to generalizations or was guided by the salience of initial syllables by randomising the structure of the language. Experiment 13 presented syllables in the same order as in Experiment 11 but with no structural relationship between the first and third syllable.

Method

Participants

10 students (who had not participated in any other experiment reported here) at the University of Warwick participated for £1.

Materials and design

This experiment is analogous to Experiment 8, which removed the structure of the language, but maintained the position of syllables within words. I used the same training stimuli as Experiment 8, with the exception that 25 ms pauses between words were added in the speech stream during training. The speech stream was composed of 27 words composed of three syllables, such that three syllables always occurred in the first position in the word, three syllables always occurred word-medially, and the remaining three syllables always occurred word-finally. Syllables occurred in the same position as in Peña *et al.*'s original experiments, thus I randomized structure, but not phonology in this Experiment. As in Experiment 8, no initial syllable began consecutive words, which was also the case for medial and final syllables. The test stimuli were composed of 36 forced-choice pairs: each pair contained one of the 36 rule-words that could be

generated (3 initial position syllables x 3 final position syllables x 4 end-item syllables in the new medial position) versus their part-word counterparts. There were 4 (6 – 2) end-item syllables in medial position because rule-words containing a repetition of the first or last syllable.

Procedure

The training and testing procedure were identical to that for Experiment 11 in every other way.

Results

We found a significant preference for rule-words over part-words (see Figure 21). Participants selected rule-words over part-words with a mean of 22.9 (63%) responses from 36, which was significantly greater than chance, $t(9) = -3.64$, $p < .01$. In addition, there was no bias for choosing rule-words against XB_iA_j part-words as opposed to rule-words versus B_iA_jX part-words, $t(9) = -.88$, $p = .40$.

Discussion

We found that participants generalized to rule-words even when there was no structure in the language. In Experiment 12, too, there was a preference for rule-words even when the position of syllables was randomized. These data put together suggests that the presence of the gaps suffices in itself to promote salience of *any* first syllable as a perceptual cue to word boundary, independent of phonological properties or indeed structural organization of the words themselves. This is attested to by the absence of preference for rule-words over

different part-words – the added salience of the initial phoneme contributes additionally towards a preference for plosives in initial position. Experiment 6 found that B_iA_jX part-words generated a smaller preference for words than did XB_iA_j part-words, as they started with a plosive. This effect is overwhelmed in the current experiment by the addition of the 25 ms gap. From these results I conclude that under these specific experimental conditions it is not possible to claim that participants generalize at all, nor that they do so on the basis of algebraic computations.

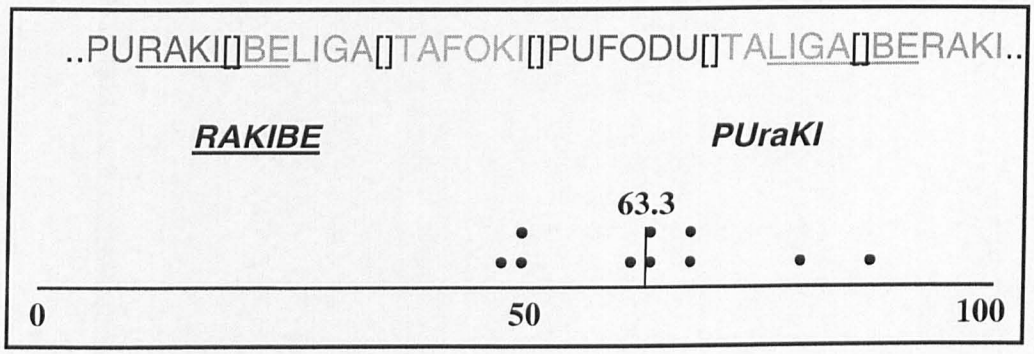

Figure 21. Experiment 13 – generalization task with gap and no structure.

General discussion

A summary of the design and the results of the experiments I have presented in this chapter is shown in Table 6. In the course of this series of experiments, I investigated whether the tasks of speech segmentation and generalization using an ALL paradigm can be separated so as to individuate two different types of mental computation, one statistical and the other algebraic, as Peña *et al.* claimed. I replicated Peña *et al.*'s experiments and ran further control experiments that do not support, at present, their theoretical segregation.

The first three experiments explored the extent to which participants exploit nonadjacent dependencies in order to individuate words in a speech stream of nonsense syllables. The overall view is that segmentation occurred on the basis of preference for plosive sounds in initial position. The remaining experiments (4-8) were concerned with the ability to generalize to rule-words using nonadjacent depedencies that included a previously encountered initial- or end-syllable in a medial position. While Experiment 9 replicated the original Peña *et al.*, Experiment 10 produced an opposite reverse effect of preference for part-words once syllable position was randomized across participants. Equally, the last three experiments (6-8), where the speech stream was interspersed with gaps between words, indicated that phonological salience of *any* first syllable was enhanced by the presence of the pauses, which in itself was sufficient to drive participants' bias for rule-words, even when the structural dependencies were eliminated.

My conclusion after meticulous observations is that there is currently no evidence from ALL experiments that people exploit nonadjacent dependencies in language learning for either segmentation or generalization. This is in accord

with a previous investigation by Newport and Aslin (2000) who, in a series of experiments, found no evidence for the learning of nonadjacent dependencies in order to segment speech. As with Peña *et al.*'s experiments, they assessed the learning of three $A_i\_B_i$ pairs when the intervening item varied among a set of three syllables.

These findings, however, pose a problem in interpreting ALL results because other experiments have indicated that nonadjacent dependencies can indeed be learned under certain conditions. Gómez (2002) found that the structure of sentences of the form $A_iX_jB_i$, where there were three different $A_i\_B_i$ pairs and sentences were presented individually, could be learned provided there was sufficient variability of X_j words. The structure was learned when 24 different *X*s were presented, but participants failed to learn when *X*s varied from sets of 2, 4, 6, or 12. Chapter 2 (see also Onnis, Christiansen, Chater, & Gómez, 2003) replicated these results and further found that structure could be learned with only one middle item, thus revealing a U-shaped curve as a function of variability. Furthermore, they found that generalization to completely novel middle items was supported only under the same conditions of no or high variability (Onnis, Gómez, Christiansen, and Chater, in preparation). Relatedly, Mintz (2002) found that participants generalised to a novel *X* in an $A_iX_jB_i$ triple in a categorisation task when there was sufficient overlap with other A_iXB_i pairs.

It seems, then, that nonadjacent dependencies can be learned in ALL tasks when there is sufficient variation and when stimuli are clearly delineated by (long) pauses. In addition, distinctions between *X*s and $A_i\_B_i$ pairs are frequently introduced in order to assist learning. In Gómez's studies, for example, *X*s have higher pitch than the $A_i\_B_i$ pairs (Newport & Aslin, 2000). One reason for the

absence of evidence for generalizations in Experiments 4 and 5 may be due to the small variability in the middle items. Another possibility for the lack of effect in these experiments is that the $A_i\_B_i$ pairs are not sufficiently distinct from the set of X_j syllables.

Another impediment to learning nonadjacent dependencies in the experiments I have presented is that concatenating words adds considerable complexity to the task of computing transitional probabilities. In Figure 22 the segmentation task used in Experiments 1 and 2 is contrasted to an ALL task with stimuli presented separately. The transitional probabilities between words (0.5) are higher than within words (0.33)[11] and this pressures for segmentation within words (Saffran, Aslin, & Newport, 1996). If variability of the middle item is increased, as seems necessary in order for generalizations to occur (Gómez, 2002), the transitional probabilities within words will drop further, but remain static for between-word transitions (Figure 23). A segmentation task version of the zero variability case (Onnis *et al.*, 2003) is not viable either. With only one X, transitional probabilities would be high everywhere; word-spanning nonadjacent dependencies ($A|X$) would have high probabilities (0.5), and would be relatively frequent in training, resulting in a seamless sequence of alternating nonadjacent dependencies (Figure 24). Natural language contains large variability of items within grammatical structures, but also lower transitional probabilities between words than within words, but these properties are difficult to simulate in small-scale ALL experiments. Until such limitations can be overcome I suggest that it is premature to conclude that statistical and algebraic computations are not performed simultaneously.

The negative results reported here and in Newport and Aslin, coupled with the positive results of Gómez and chapter 2 instruct us that the issue of what is learned in ALL paradigms cannot be settled conclusively without a thorough investigation of the interactions between experimental tasks, training procedures, and distributional properties of the input being sampled. Perhaps, then, I have failed to separate the two computational processes partly because of intrinsic experimental limitations, and partly and most importantly because the separation of computational processes is the wrong approach to the issue. If a structure like nonadjacent dependencies given very similar and comparable training material has been shown to be learned in some but not in other conditions, then the core issue is not whether it is instantiated in terms of algebraic or statistical computations, but what makes it learnable and not learnable in different conditions. In addition, are there structures that cannot be learned because of their complexity? Gómez has proposed that learners may attend to different sources of information and prefer the most statistically reliable source in order to reduce uncertainty. Specifically, whether the cognitive system focuses on bigrams, trigrams, or long-distance dependencies is largely driven by the statistical landscape. As attested by the U-shape found in chapter 2, the fact that learners fail in certain conditions, e.g. low variability of intervening items, does not entail that learners are unable to learn in other conditions.

My results are a salutary reminder that ALL experiments need careful experimental control, and my results point towards phonological preferences being of profound importance in the construction of such controls. The fact that the stimuli used are "artificial" does not mean that the building features they are

[11] If computations are independent of nonadjacent dependencies, then participants ought to segment within words rather than between words. We did not find preference for part-words over words when phoneme

composed of (in this case the phonetic features of syllables) are completely new to learners. Randomization of syllables in different positions across subjects should thus be adopted in further studies to ensure sound experimental practice even when the stimuli appear to be either perceptually or conceptually artificial.

My work is reminiscent of the debate between Johnstone and Shanks (1999) and Meulemans and Van der Linden (1997). The latter presented evidence for the same separation invoked by Peña *et al.* between a mechanism based on knowledge of chunks of letters in the training strings and the other based on algebraic rules. They constructed a measure of chunk strength for their stimuli, creating four groups of string items: grammatical and associated (GA), nongrammatical and associated (NGA), grammatical and nonassociated (GNA), and nongrammatical and nonassociated (NGNA). Associated test strings contained bigrams and trigrams that occurred significantly more frequently than in nonassociated strings, as measured by the associative chunk strength metric. When participants were exposed to few items at training (their Experiments 1A and 2A) they classified associated test items more often as grammatical than nonassociated ones. Conversely, when most of the grammatical items were presented in the learning phase, Meulemans and Van der Linden claimed that only an effect of rule abstraction was observed. In reappraising these conclusions, Johnstone and Shanks (1999) argued that not only were learners sensitive to chunk frequency, but they also gained information during training about the legal locations of chunks within training strings. They demonstrated that participants classified test strings as ungrammatical not because they violated the rules of the grammar but because they contained chunks in *novel* locations with respect to training strings. For instance, the training set contained

preference was controlled for, due perhaps to the small differences in transitional probability.

the trigram VXR occurring 10 times in only 2 of the 5 possible locations (MXRM*VXR*, M*VXR*, M*VXR*VVV, M*VXR*MXT, M*VXR*MXR, M*VXR*VMT, M*VXR*VV, M*VXR*V, M*VXR*M, and M*VXR*VVM), string length spanning from three to nine letters. It turned out that the same chunk appeared in a new position, namely as fourth trigram in four out of eight nongrammatical test strings (MXR*VXR*M, MXR*VXR*V, VMR*VXR*M, VMR*VXR*V). This chunk positional information coupled with chunk frequency could account for the highest proportion of the variance in multiple regression tests, thus explaining away an effect of grammaticality due to abstraction of the rules of the grammar.

In general, distinguishing empirically between algebraic rules and other forms of knowledge in ALL paradigms remains an elusive problem. In the first place this is because the exact nature of rule-based knowledge has been left rather vague, despite strenuous argumentations have been put forward for rule-based learning (see for instance the discussion between Marcus & Berent, 2003 and Seidenberg, MacDonald, & Saffran, 2003). The fact is that rules are invoked whenever some structure does not seem to be learnable from other sources of information. The idea of algebraic rule entails the formation of a higher-level abstract mental representation that describes the states (nodes) of the finite-state grammar that generated the stimuli. Generalization tasks are often taken as a test-bed for algebraic computation because they require abstraction to novel stimuli. Such abstraction can in fact be couched both in algebraic terms and in statistical terms. For instance, Marcus *et al.* (1999) exposed seven-month-old infants to seamless speech strings containing one of two word patterns, *ABA* (*de-li-de* and *wi-di-wi*) or *ABB* (*wi-di-di* and *de-li-li*). While training and test strings contained the same pattern, test strings were instantiated with new words (*ba-po-ba* or *ba-*

po-po). Using a preferential-looking procedure, infants showed familiarity for strings belonging to the training pattern but not for strings with a different pattern, despite the change in the strings' surface form.

Marcus and colleagues interpreted these powerful abstraction abilities as incontrovertible evidence for the existence of algebraic computation. However, abstraction at test could be based on recognizing the perceptual similarity of the physical stimuli, for instance noting that instances of the pattern *ABB* contain two physically identical items. Brooks and Vokey (1991) argued that repetition patterns could be a sufficient indication of the goodness of a test item: in this respect MXVV could be a good match for HJLL without appeal to algebraic rules. In fact, Gómez *et al.* (2000) found abstraction beyond specific word order only for grammars that contained repeating elements. This result is suggestive of potentially different levels of abstraction. Pattern-abstraction operates through comparison over physical stimuli. Conversely, acquiring linguistic representations such as Noun-Verb-Noun patterns (*John loves Mary*), requires a knowledge that is category-based, i.e. it involves generalizations that are abstract and perceptually unbound (*John* and *Mary* are orthogonal instantiations of the category *Noun*), as well as positionally unbound, at least partially (*John* and *Mary* can be swapped in the chain to obtain *Mary loves John*; see Gómez & Gerken, 2000). In addition, Christiansen, Conway, and Curtin (2000) successfully simulated the experiment by Marcus *et al.* using connectionist models that learn by simple associative mechanisms, illustrating the precariousness of separating algebraic from statistical computations.

Redington and Chater (2002) have argued that evidence for abstraction, i.e. surface-independent knowledge, does not imply that knowledge is also rule-

based, as these concepts are orthogonal. All sources of information, including positional, and surface-based, can be instantiated in a symbolic rule. Indeed, my results could all be couched in symbolic terms. For instance, learners might have internalised the following rule, based on their acquired knowledge of the English lexicon: "A syllable that begins with the sounds /p/, /b/, or /t/ appears in a word in initial position". Even strings that were constructed without nonadjacent dependency structure could still be represented in the symbolic rule: "/pu/, /be/, and /ta/ appear in first position while /ki/, /ga/, and /du/ appear in last position". Indeed, this is exactly the instruction I wrote in the computer script that generated the stimuli for my experiment.

Framing my results within the recent ALL literature, it appears that there is a cascade of potential cues that learners might pick up on in order to detect structure: conditional probabilities, nonadjacent dependencies, positional information, similarity with previously seen items, gaps between elements, etc. My experiments have contributed by elucidating the role of phonological sensitivity, a cue that so far has been underplayed in ALL studies but has been shown to have a potentially vast role in language acquisition (Cassidy & Kelly, 1989, 1991; Kelly, 1992; Monaghan, Chater, & Christiansen, submitted). In my view casting the study of human learning in terms of rules versus statistics may be an ill-posited research program far less central to understanding the human mind than, for instance a) investigating the training and test conditions in which learning takes place, and verifying whether learning transfers across modalities; b) determining whether infants and adults learn the same structures in comparable conditions; or c) in the face of multiple cues, determining whether learners integrate them, discard the less reliable ones, or choose one in a winner-

takes-it-all fashion (see, e.g., Saffran, Newport, & Aslin, 1996). Overall, there is wide and growing evidence that language phenomena are probabilistic in nature at all levels of analysis, and what needs to be tackled theoretically is how to capture this probabilistic nature (Bod *et al.*, 2003).

The remarkable finding from my studies is that, even when there is no statistical structure in the language, participants demonstrate a stable preference for certain speech sounds occurring in given positions. And the addition of short gaps between words affords salience of any initial syllable as a reliable word initial candidate. Phonological preferences impact both segmentation and generalization tasks in ALL, and, in the series of experiments presented here, obscure any statistical or algebraic computations on the speech signal that might take place. The surprising conclusion from Peña *et al.* (2002) that statistical and algebraic processes are distinct in language learning proves to be premature.

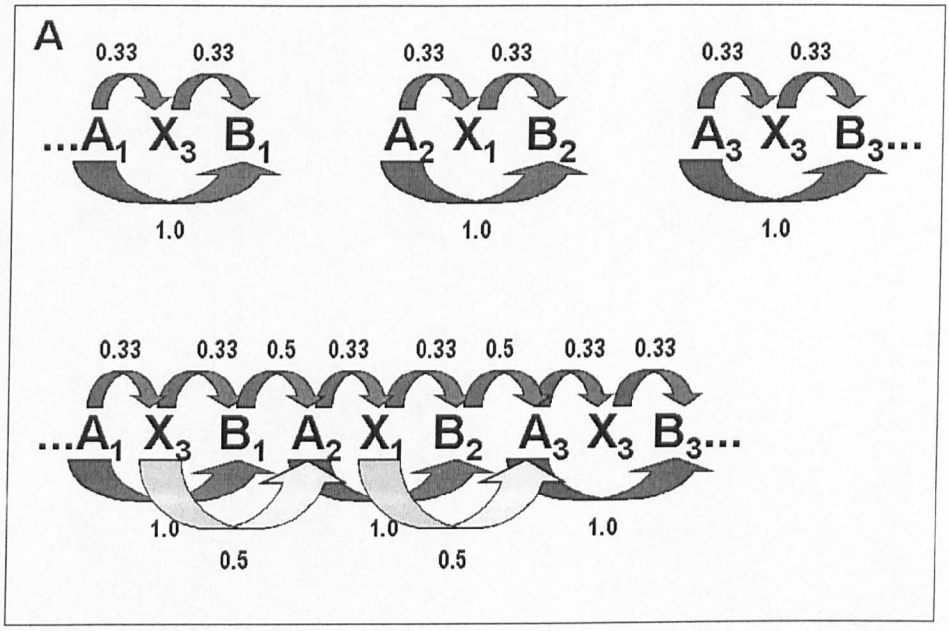

Figure 22. Comparison between a traditional ALL task (above) and the segmentation task used by Peña *et al.* (below).

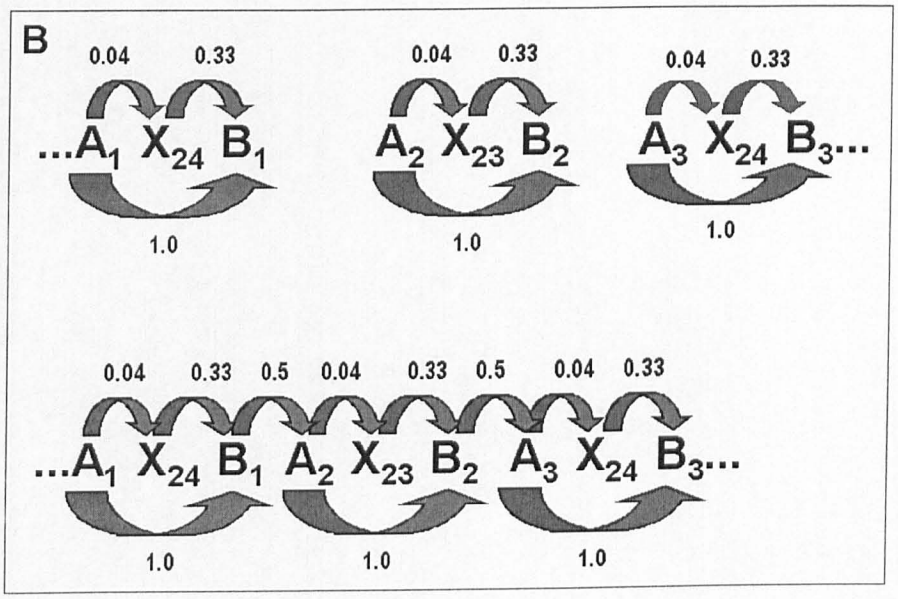

Figure 23. Comparison between the ALL task used by Gómez (2002) with large variability of middle items (above) and a hypothetical mirror segmentation task (below), where low-transitional probabilities between the *A*s and the *X*s would lead to wrong segmentation.

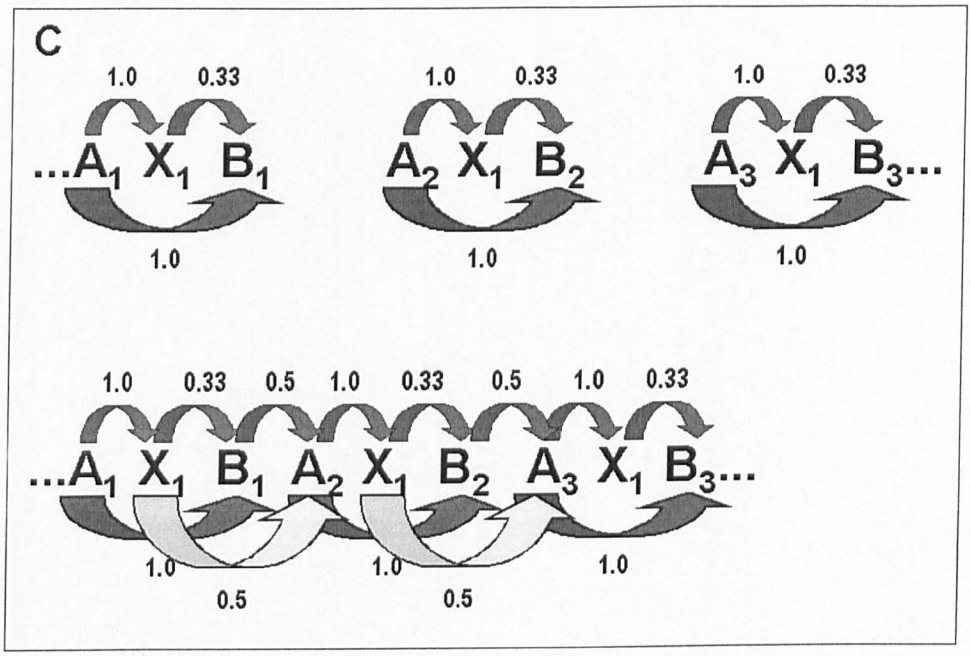

Figure 24. Comparison between the ALL task used In chapter 2 with no variability of middle items (above) and a hypothetical mirror segmentation task (below), where unwanted nonadjacent dependencies between the *X*s and the *A*s having relatively high conditional probabilities would lead to an impossible task

Chapter 7

Recovery from overgeneralizations in language acquisition

Natural languages are most often characterized as a combination of rule-based generalization and lexical idiosyncrasy. The English past tense is a familiar case, in which the irregular form *went* replaces the expected +ed construction **goed.* Baker (1979) notes that this is a relatively benign example for learners, since irregular forms are frequently encountered in the course of their linguistic experience. The experience of the form *went* may block **goed*, if the learner assumes that verbs typically have a single past tense form – thus, an observed alternative form can serve as evidence that an absent regular form is not allowed in the language (e.g. the Competition model, MacWhinney, 1989). Much more troubling are cases where an apparently legal construction is idiosyncratically absent, without any alternative. The dative shift in English is a well-documented example:

(1) *John gave/donated a book to the library*

(2) *John gave/*donated the library a book*

In such cases we can think of linguistic rules as being quasi-regular: they license the combination and production of *some* members of syntactic categories, but not others. The difficulty of learning such idiosyncratic absences from partial input and without negative evidence (as is the case with natural language) has become notorious in the language acquisition literature. In particular, given that only a finite set of sentences is ever heard, out of the infinite set of possible sentences in

a natural language, it is clear that mere absence of a linguistic form cannot be directly used as evidence that the form is not allowed. Yet, such 'holes' are clearly specific to particular natural languages, and hence cannot be explained by adversion to innate linguistic principles. This problem has been viewed as so severe that it has been labeled Baker's *paradox*; and viewed as raising *logical* problems for the theory of language acquisition (e.g., Baker & McCarthy, 1981)[12]. The approach I adopt here is to apply a general principle of learning to explain how linguistic idiosyncracies can be acquired. Note that the mechanism must be sufficiently flexible to capture the huge range of idiosyncrasies across a vast range of linguistic contexts. Moreover, the existence of such a mechanism is required, I contend, to explain the existence of idiosyncracies in language evolution: idiosyncrasies could not have emerged or survived in its absence, as they would have been winnowed out by learning failures by successive linguistic generations. In this respect, Baker's paradox raises a secondary paradox for language evolution, which is dealt with in the next chapter. The puzzle of how language acquisition processes can capture what appear to be idiosyncratic 'holes' in the language also raises the puzzle of how difficult-to-acquire linguistic patterns emerge and are transmitted in the development of languages. Note that, on pain of circularity, whatever learning mechanisms are responsible for learning such

[12] Many writers have argued that the general problem of language acquisition inevitably necessitates innate language-learning modules: "no known 'general learning' mechanism can acquire a natural language solely on the basis of positive or negative evidence, and the prospects of finding any such domain-independent device seem rather dim" (Hauser *et al.*, 2002: 1577. See also Chomsky, 1957; Pinker, 1989). Gold (1967) has shown that language identification in the limit is impossible for a broad class of formal languages. By contrast, Horning (1969) has shown that grammatical inference is in a probabilistic sense, for languages generated by stochastic context free grammars. More recently, Chater and Vitányi (2001) have shown that such inference is possible for any computable language, including, a fortiori, any grammars involving context sensitivity and/or transformations, if the goal is (arbitrarily close) agreement between the learner's language with the target language. The method that underpins Chater and Vitányi's theoretical result is practically implemented in the simulations described here – the learner seeks the simplest description of the corpus it has received.

idiosyncracies must pre-date the emergence of such idiosyncracies. That is, we cannot view the idiosyncratic nature of language as a stable environment to which biological basis for language acquisition adapted – because without relevant prior learning mechanisms already established, language could not have developed with such idiosyncracies in the first place.

Having considered how a cognitive system might learn to detect structure and generalise from experience in previous chapters, in this chapter I consider the question of learning idiosyncracies by recovering from linguistic overregularisations. Firstly, I begin by outlining why they constitute such an apparently difficult learning problem. Secondly, I summarise a small number of putative mechanisms that have been put forward in the literature. Thirdly, I present a model that is able to learn quasi-regular structures in a rudimentary language from positive evidence alone, using a very general learning principle: simplicity. The model learns by creating competing hypothetical grammars to fit the language to which it has been exposed, and choosing the simplest. As an explicit metric for simplicity I use Minimum Description Length (MDL), a mathematical idea grounded in Kolmogorov complexity theory (Li & Vitányi, 1997). In acquiring quasi-regular language structures, this model specifically addresses the acquisition problem.

Baker's Paradox and linguistic quasi-productivity

A mainstay of linguistic analysis has been that human languages are composed of a limited number of basic units (features, segments, syllables, morphemes, words, phrases, clauses, etc.) that can be combined by a small number of generative rules

to create larger units. Postulating the existence of recursive rules allows for an infinite number of sentences to be created. This generativity goes well beyond theoretical linguistic description, as it is typically taken to be embodied in the psychological mechanisms responsible for acquiring and representing linguistic rules and units.

Although the capacity to generalize from a limited set of examples to novel instances is an uncontroversial aspect of the human cognition, a puzzle that has attracted linguists is that natural languages, although productive, are never fully regular. There appear to be finely-tuned lexical and syntactic selectional constraints that native speakers are aware of. Expected regular structures may either be replaced (e.g. *went* for **goed*) or they may be disallowed completely. These semi-productive structures may be seen as a special case of irregularity where the irregular form is absent, i.e. there seems to be an unfilled slot that constrains open-ended productivity. Consider, for instance, a transformational rule such as *to be* Deletion (after Baker, 1979):

(3) X – to be – Y

→ *X, 0, Y*

(4) The baby seems/appears to be happy

(5) The baby seems/appears happy

(6) The baby seems to be sleeping

(7) The baby happens to be sleepy

*(8) *The baby seems/appears sleeping*

*(9) *The baby happens sleepy*

On the basis of positive evidence positing the transformational rule in (3) is misleading with regard to the perfectly plausible but ungrammatical predictions that it gives about sentences (8) and (9). Such 'unfilled slots' cannot be accounted for by the general rule. Similarly, consider the lexical constraints on the collocations between, for instance, adjective and noun below:

(10) strong/ high/*stiff winds

(11) strong /*high/*stiff currents

(12) strong/*high/stiff breeze

Quasi-productivities are ubiquitous in the lexicon and it has been proposed that they constitute a considerable portion of syntax as well (for a discussion of the vast range of syntactic idiosyncrasies including wh-movement and subjacency, see Culicover, 1999). In standard generative grammar these 'syntactic nuts' have traditionally been disregarded as the 'periphery' of the language system, where the 'core' is a set of general fully regular principles requiring a minimum of stipulation. Most syntactic constructions, however, are subject to varying degrees of lexical idiosyncrasy. Consider another familiar example, the constraints on the Dative shift transformation:

(13) $NP_1 - V - NP_2 -$ *to* NP_3

$\rightarrow NP_1, V, NP_3, NP_2$ *(optional)*

(14) We sent the book to George

(15) We sent George the book

(16) We reported the accident to the police

*(17) *We reported the police the accident*

Indeed, as Culicover (1999), and others within the general movement of construction grammar (Goldberg, 2003), have argued, such idiosyncracies may be so ubiquitous that the 'periphery' of standard linguistic theory may encroach deep into the 'core,' of standard linguistic theory – so much so, indeed, that explanatory principles and learning mechanisms required to deal with the periphery might even deal with the core as a limiting case.

To see why the presence of semi-productive regularities represent a particularly difficult learning problem, I now consider arguments concerning language learnability and the contribution of innate linguistic constraints.

The logical problem of language acquisition

At a general level, the so-called logical problem of language acquisition is that learning a language from experience alone is impossible because linguistic experience is too incomplete and contradictory. In the first place, a learner observes only a limited set of the infinite number of utterances in his/her language. From this, he/she must distinguish a certain set of 'grammatical' utterances among all the other utterances that he has never heard and may never produce. The problem is particularly acute when considering the case of quasi-productivities, which yield Baker's Paradox (also known as the Projection

problem) after Baker (1979). Baker noted that quasi-productive regularities such as those above pose a genuine puzzle for any account of language acquisition. This is principally because the unfilled slots they create in the language occur *within* the space of allowable sentences and nonetheless are somehow blocked by language learners. A crucial tenet of the logical problem is that indirect negative evidence in the form of absence is not sufficient to constrain the learner's hypotheses about the correct grammar, because there are many linguistic sentences that a learner has never heard but are nonetheless grammatical (Pinker, 1994). There are therefore many hypothesis grammars that would be consistent with the positive data available. It is suggested that such a hard learning problem necessitates the existence of powerful innate linguistic tools. Since the literature has polarised around the acquisition of verbs' argument structure, I focus on such examples throughout this and the next chapter. Before I dwell on the simplicity model, I summarise two popular accounts that start from different assumptions, the semantic bootstrapping model, and the construction grammar approach.

Learning Argument Structure: semantic bootstrapping

One proposal involving innate linguistic rules comes from Pinker (1984). Pinker has proposed the Semantic Bootstrapping Hypothesis to account for the acquisition of Verb Argument Structure, whose main point is that the productivity of lexical rules is governed by semantic criteria determining which verbs they can apply to. Pinker distinguishes 6 broad semantic classes of argument structure:

Simple transitives

Datives (I will tell/*shout you the message)

Locatives (I poured/*filled glass into the water, I filled the glass with water)

Passives (*Amy is resembled by Sue)

Resultatives (Betty wiped the table clean)

Causatives (I broke the glass/the glass broke, I cut the bread/*the bread cut)

Each broad semantic class is associated with characteristic semantic properties, or thematic cores. For instance, the transitive construction has the following semantics associated to it:

X acts on Y

The problem with learning is that within each broad class verbs behave differently. Some may take two different syntactic structures (these verbs are said to alternate), whereas others are restricted to only one of them. Let us take the dative alternation as an example. The dative alternation has two forms:

1- the ditransitive form NPx-V-NPy-NPz (e.g. I sent Mary a package)
2- the prepositional form NPx-V-NPz-to/for-NPy (I sent a package to Mary)

where *NPy* represents the recipient or goal. Not all dative verbs alternate, and not all those that alternate do so in all contexts. Pinker distinguishes three narrower classes:

1) verbs that alternate (i.e. accept (a) and (b) above): give, bring, offer, send, build, promise, make, get, buy, take, throw, leave, forward, refer, allocate, guarantee, allot, award, grant, reserve.

Within this subclass a semantic restriction applies. We do not say *I sent the border a package*. Pinker argues that the ditransitive and the prepositional forms have two underlying semantic representations, respectively:

1- to cause Y to have X (double-object dative)

2- to cause X to get to Y (prepositional dative)

2) verbs that accept only the ditransitive form: ask, envy, bet, refuse, charge, forgive, spare, lend, teach, cost, deny, fine, tell, show.

3) verbs that accept only the prepositional form: carry, supply, recommend, describe, stir, taste, demonstrate, choose, donate, explain, report, recite, construct, deliver, dictate, contribute, reply, present, design, shout.

4) verbs that accept only the full prepositional form: *credit, entrust, reward, present, honour* (usually "with").

So for Pinker "membership in a broad conflation class is only a necessary condition for a verb to alternate" (p.103). The meaning component added by an argument structure cannot in itself explain so-called negative exceptions such as the following:

(18) John took Mary the ball

*(19) *John carried Mary the ball*

What determines the alternation is the membership of each verb to a small set of narrow conflation classes, which are sensitive to subtle semantic distinctions. Crucially, the correct association between a verb's semantic structure and an argument structure is carried out via linking rules, which are innate. Children learn to avoid generalisations by learning more and more accurate meanings for more and more verbs. Once the child has correctly identified the verb's meaning generalisation errors should disappear.

Several critiques have been levelled at Pinker, notably because his proposal does not seem to have empirical support. If narrow conflation classes are learnt lately by children, one would expect children to be overproductive earlier than 3-4 years of age (Bowerman, 1990). In addition, Slobin (1998) argues that innate rules are too general to constrain all languages across different forms over historical time. And there is much variation across languages as to what verbs take which argument structure. It seems that for Pinker language is a completely logical system and that the child only needs to discover progressively this system by application of innate linking rules. The exact functioning of such rules remains unclear and yields little explanatory power. Pinker discards a priori the existence of indirect negative evidence in the form of non-occurrence as being a surrogate for negative evidence. In his words, "there is always an infinity of sentences that [the child] hasn't heard that are grammatical" (page 14), so indirect negative evidence is simply a restatement of the learning problem.

Learning Argument Structure: Construction Grammar

A different view on how children learn verbs' argument structure is provided by Goldberg (1995), who proposes a construction grammar approach. The semantics of argument structure cannot be associated completely to a specific verb because verbs usually appear in multiple argument structures. Also, many verbs share the same argument structure. At the same time, there appear to be regularities between form and meaning of an argument. So the construction, SUBJ–Verb – OBJ1 –OBJ2 carries the meaning of transfer. In Construction Grammar, C is a construction iff it is a pairing of form and function such that some aspect of the form or the function is not strictly predictable from the component parts of C. Constructional meaning in learning arises from so-called "light verbs" (do, make, take, go, give, put, find). These are highly frequent and learned early. In a first phase, AS is initially associated on an item-by-item basis. As vocabulary increases, abstract constructions emerge. This is in line with most work done by Tomasello and other researchers recently. AS emerges from being associated with light verbs. In the Semantic Bootstrapping Hypothesis (Pinker, 1989) meaning is predictable given a complete lexical specification of a verb's meaning and innate linking rule. Syntax is highly abstract, while it is the lexicon that contains all information. In Construction Grammar on the contrary, it is possible to have a generalisation such as *She sneezed the foam off the cappuccino* although the verb *sneeze* is intransitive. This is because the argument structure SUBJ-Verb-OBJ OBLlocative captures the meaning of caused motion. There is no need to have an

entry in the lexicon with a special transitive meaning for *sneeze*. For the inseparability of syntax and lexicon see also Bates & Goodman (1997).

Other researchers have proposed that children learn argument structure by exploiting both semantic and syntactic cues (Gleitman, 1990). Allen (1997) developed a connectionist network that learned argument structure using both syntactic and semantic cues extracted from a sample of the CHILDES database.

Learning Argument Structure from non-occurrence

The paradox raised by Baker is that even postulating a Universal Grammar that restricts the search space for potential grammars does not solve this particular problem, since unfilled slots are highly idiosyncratic across languages. I contend that because these constructions cannot be derived from universal principles, they must be determined by the learner on the basis of exposure to the language, thus providing a solution to Baker's paradox. Although semantic knowledge may help the learner, this apparently intractable computational problem will not disappear in the face of simple appeal to semantics. For instance, we have seen that transitive and intransitive verbs may be distinguished by virtue of the fact that transitive verbs refer to sequences involving both agents and patients, whilst intransitives involve only agents:

(20) John broke the cup

(21) The cup broke

(22) John kissed Mary

*(23) *Mary kissed*

However, Bowerman (1996) has noted that it can be misleading to predict syntactic behaviour from semantics, for instance *donate* and *give* in examples (1) and (2) have similar semantics but *donate* does not allow for dative shift. It is worth noting that younger speakers of English will often fail to judge the phrase *John donated the library a book* as ungrammatical. This may be an example of regularization, but this does not weaken the argument. Consider also:

(24) John waved Mary goodbye

(25) John waved goodbye to Mary

*(26) *John said Mary hallo*

(27) John said hallo to Mary

or, again from Baker:

(28) It is likely that John will come

(29) It is possible that John will come

(30) John is likely to come

*(31) *John is possible to come*

Hence, I argue that some degree of arbitrariness must be accounted for in quasi-regular constructions (see also Culicover, 1999, on the case for at least partial independence of syntax from semantics in the case of unfilled slots). If idiosyncrasy is to be found at the core of grammar and can neither be accounted

for by universal principles nor semantically determined completely, it must be learnable from experience, possibly from a distributional analysis of the input.

Causative alternations in child-directed speech

Suppose we have a language in which verbs belong to three distinct classes (*V1*, *V2*, *V3*). Each class is related to two syntactic contexts (*C1*, *C2*). One class of verbs (*V1*) appears in both contexts. Two other classes of verbs (*V2* and *V3*) occur in one context only. We can produce a simple table to visualize the alternation:

	C1	C2
V1	1	1
V2	0	1
V3	1	0

Table 7. Alternating and non-alternating verbs across contexts.

The causative alternation in English is of this kind. Verbs like *break* behave both transitively (*I broke the vase*) and intransitively (*The vase broke*), whereas verbs like *disappear* behave only intransitively (*The rabbit disappeared* is allowed; but **I disappeared the rabbit* is not) and verbs like *cut* are found only in transitive contexts (**The bread cuts* is not allowed). An analysis of CHILDES revealed that verbs in child-directed speech fit the pattern of the above idealization: a number of verbs are exclusively transitive or intransitive (see Table 8).

Children eventually generalize the structures of the language they are exposed to. A typical generalization occurs when children say *Don't you fall me*

down (Bowerman, 1982; Lord, 1979). This is an overgeneralized use of a non-causative verb as causative. In the causative construction, some verbs like *break* can be used both transitively with a semantic element of cause (*I broke the vase*) and intransitively (*the vase broke*). Verbs like *break* alternate between two constructions. However, *fall* can only be used intransitively, and *hear* only transitively. The acquisition of verbs' argument structure seems particularly complicated as the way verbs behave syntactically is largely arbitrary. Semantically similar verbs like *say* and *tell*, or *give* and *donate* allow for different constructions.

Bowerman (1982) and Lord (1979) recorded a total of 100 different cases in which two-argument verbs are used with three arguments (e.g. *You can drink me the milk*). The developmental literature suggests that when children acquire a new verb they use it productively in both constructions, without specific directional bias (Lord, 1979). It is also worth noting that alternations can be theoretically distinguished from other forms of irregularization like the irregular past tense. In the case of *goed-went* for example, recovery from the overgeneralized form **goed* can be accounted for by directly invoking a competition strategy (MacWhinney, 1987): as the number of *went* in the input increases, it will win over the irregularised form *goed*, which has 0 frequency in the input. Alternations are interesting theoretically in that the competition model does not seem applicable for these. The overgeneralized form does not have an irregular alternative: there is simply a "hole" in the language. This argument was raised by Baker in his distinction between benign exceptions (like the past tense) and truly problematic alternations like the ones I consider here (Baker 1979).

For the purpose of showing how such problematic irregularities can be learnt using a simplicity principle, I take the causative alternation described above as a working example. I extracted verb frequencies from the CHILDES Database. CHILDES contains a total of nearly ten million words of child-directed speech. Because I am interested in showing that the input the child receives is rich enough for recovery of overgeneralization by induction, only the adult speech in the corpus was selected and analysed.

	Verb	Transitive occurrences	Intransitive occurrences
	bounce	75	117
	break	1251	268
	burn	86	60
	close	855	56
	freeze	18	61
	grow	59	330
Category	move	966	560
V1	open	1590	232
	pop	104	153
	rip	139	9
	roll	405	164
	shake	147	26
	slide	65	120
	swing	38	96
	tear	167	20
	turn	2690	600
	arrive	0	41
	come	0	18437
	dance	0	370
Category	die	0	141
V2	disappear	0	73
	fall	0	2945
	go	0	65193
	rise	0	14
	run	0	1569
	stay	0	1413
	bring	3028	0
	cut	1315	0
	drop	640	0
Category	kill	120	0
V3	lift	392	0
	push	1609	0
	put	27154	0
	raise	25	0
	take	9724	0
	throw	2090	0

Table 8. Verbs in child-directed speech occurring in transitive and intransitive contexts pooled from the CHILDES English sub-corpora (MacWhinney, 2000).

Simplicity and Language

The simplicity principle (Chater, 1996) states that in choosing among potential models of finite data, there is a general tendency to seek simpler models over complex ones and optimize the trade-off between model complexity and accuracy of model's description (i.e. fit) to the training data. Complexity is thus defined as:

$$C = C(\text{model}) + C(\text{data}|\text{model})$$

The favoured model of any finite set of data will be that which minimizes this term.

In order to compare different grammars we need a measure of simplicity and a "common currency" for measuring both the model complexity and the error term complexity. Fortunately this is possible by viewing grammar induction as a means of *encoding* the linguistic input; the grammatical organization chosen (the "knowledge" of the language) is that which allows the simplest encoding of the input. A tradition within mathematics and computer science, Kolmogorov complexity, shows that the simplest encoding of an object can be identified with the shortest program that regenerates the object (Li & Vitanyi, 1997).

Every sentence generated from a lexicon of n words may be coded into a binary sequence. The length of a message refers to a binary string description of the message in an arbitrary universal programming language. The binary string can be seen as a series of binary decisions needed to specify the message; smaller lengths correspond to simpler messages. The brevity of an input A_i is associated

to its probability $P(A_i)$ of occurrence. Shannon's (1948) noiseless coding theorem specifies that:

$$Length=Log_2[1/P(A_i)]$$

More probable events are therefore given shorter codes. Li & Vitanyi (1997) have shown that the length K(x) of the shortest program generating an object x is also related to its probability Q(x) by the following *coding theorem*:

$$K(x)=log_2[1/Q(x)]$$

Finally, the *invariance theorem* (Li & Vitanyi, 1997) assures that the shortest description of any object is *invariant* (up to a constant) between different universal languages, thus granting a measure of simplicity that is independent of the data and of the programming language used to encode the data. The above formalizations allow us to replace "Complexity" with "Length" and state that "the best theory to infer from a set of data is the one which minimizes the length of the theory and the length of the data when encoded using the theory as a predictor for the data" (Quinlan and Rivest, 1989; Rissanen, 1989).

Modeling language learning with simplicity

In any study of grammar induction, and in particular in the simplicity framework, it is crucial to see a grammar as a *hypothesis about the data*. The best hypothesis is the one that compresses the data maximally, so we can also think of a grammar as compression of the data. We can see the achievement of adult linguistic

competence as a process of building different hypotheses about the language in order to achieve optimum compression. The essence of compression is to provide a shorter encoding of the data, enabling generalizations and correct predictions. Alternations are particularly informative about the possibility of a cognitive system to capture dependencies from limited data. If linguistic structures were completely regular, then generalizing from a few data would be easy. But as alternations are quasi-regular, meaning there are exceptions to their regularity, a learner must capture fine dependencies in order to generalize whilst avoiding overgeneralizations.

The issue is to choose the candidate model of the right complexity to describe the corpus data, as stated by the simplicity principle. We can compare different hypotheses (grammars) at different stages of learning and choose, for each stage, the one that minimizes the sum of the grammar-encoding-length and the data-encoding-length. In the following section I compare data compression of corpora by two similar models. The difference between them is that one posits a completely regular rule, whilst the other posits a regular rule and some exceptions to it. We can think of the second model as having 'invested' in exceptions. Each exception initially produces less compression overall, since the exceptions cost some bits to specify. However, each exception shortens the code-length for each item in the corpus, and the second model thereby 'recoups' its investment over time.

The Models

This approach to language acquisition does not focus on how learning occurs. Rather, these simulations run several models concurrently to show that the rate of increase of code-length differs between structures. This section describes the

structure of two hypotheses (grammars); the first gives rise to overgeneralization phenomena whilst the second does not. These were designed in conjunction with a very simple artificial language, which was subsequently used to test the models. A brief outline of the language is given here to facilitate the description of the model. A more detailed consideration of how the artificial language relates to data from corpora of child-directed speech is given below.

The artificial language used consists of two syntactic categories. These can be thought of crudely as nouns and verbs. They can be combined to form two-word sentences. Sentences may be of the form *NV* or *VN*. Forms *NN* and *VV* are disallowed. In addition, a number of sentences are disallowed. Let us imagine that there are four nouns (n_1-n_4) and four verbs (v_1-v_4) in the language, and that v_4 is blocked in the sentence final position. From this it follows that four sentences are disallowed: each of the four nouns in combination with v_4 in an NV-type sentence.

Each model is comprised of 4 elements: word-level categories, sentence-level categories, exceptions, and code-length. Both models described here contain two word-level categories, comprising nouns and verbs and two sentence-level categories comprising the two sentence types (*NV* and *VN*). The exceptions category discretely specified all the disallowed sentences. In the first model this was an empty set. The code-length specified length of code, in bits, that would be needed to specify models just described and the corpus data given the model structure. The code-length for each sentence in the corpus is consequent on the model structure.

Calculating Code-Length for each element

The length of code necessary to specify any object, *i*, is given by:

$$\text{Bits}(i)=\text{Log}_2(1/p_i) \quad [1]$$

where p_i is the probability of object *i*. In many cases described below, p_i can be thought of as choosing one of I options. Where this is the case,

$$\text{Bits}(i)=\text{Log}_2 I \quad [2]$$

This section describes how this formula is applied to calculate the code-length for each section of the model and for the data given the model.

If a language contains *r* word types and *n* syntactic categories, then the probability of specifying one distribution of word types into categories is the inverse of the number of ways in which r word types can be distributed between n categories, assuming no empty sets. This is given by:

$$\text{Distributions}(r, n)= \sum_{v=0}^{n}(-1)^v \frac{(n-v)^r}{(n-v)!\,v!} \quad [3]$$

The codelength for the word-level element is therefore:

Word-level bits(r, n)

$$=\text{Log}_2 \sum_{v=0}^{n}(-1)^v \frac{(n-v)^r}{(n-v)!\,v!} \quad [4]$$

Specifying a particular sentence-level rule (e.g. that a sentence may be of the form *NV*) is a function of the probability of that sentence type given the number of categories specified in the word-level element. Given that in the artificial language sentences only ever contain two words, there are four sentence types possible from two syntactic categories (*NN*, *NV*, *VN*, *VV*). The probability of any sentence type (e.g. *NV*) is therefore 1/4. When this has been specified, the probability any remaining sentence type (e.g. *VN*) is 1/3. The code-length for specifying two sentence types is therefore:

$$\text{Sentence-level bits}=\text{Log}_2(4)+\text{Log}_2(3) \quad [5]$$

Specifying the cost of an exception is the same as specifying the cost of a sentence. This is done by specifying the cost, in bits, of the first word based on the probability of its occurrence, and the cost of the second word in the same way. The probability of a word's occurrence is the inverse of the total number of possible words. The term to specify the first word in any sentence is therefore:

$$\text{Bits}(i1)=\text{Log}_2(\text{Tw}-T_{e1}) \quad [6]$$

where Bits($i1$) is the bits required to specify word i in the first position, T_w is the total number of word types in the language and T_{e1}, is the total number of words blocked in the sentence initial position as listed in the exceptions category.
The first word specifies which sentence type is being used. The pool of possible words from which the second word must come is therefore reduced to the size of the sentence final category as defined by the sentence type. For example, if the

first word in a sentence is a noun, the sentence type must be NV and the second word must therefore be from the category V. The term to specify the second word in a sentence is therefore:

$$Bits(j2)=Log_2 (T_{wc}-T_{e2|1}) \qquad [7]$$

where Bit($j2$) is the number of bits required to specify word j in the second position, T_{wc} is the total number of word types in category c, and $T_{e2|1}$ is the total number of words specified in the exceptions element as blocked in position two given the word in position 1. The number of bits for specifying any sentence i,j is simply:

$$\text{sentence bits}_{i,j}=Bits(i1)+Bits(j2) \qquad [8]$$

Specifying the code length for each exception is the same as specifying code length for a sentence *given the existing exceptions*. Each exception in a list of exceptions therefore requires slightly fewer bits to code than its predecessor.

It is important to note that these models code corpus data in batch mode – the order in which sentences are coded is not taken into account. A more psychologically realistic (i.e. incremental) algorithm might make use of the fact that frequently occurring words have a higher probability of occurrence and therefore cost less to code. A refined and incremental model is presented in chapter 8 to account for the transmission of irregular languages over generations of simulated learners.

Simulating recovery from overgeneralization with an artificial language

The models described above were implemented in a computer program. They were then exposed to successively large corpora of sentences from an artificial language, which reflected the structure of the transitive/intransitive alternation phenomena found in the CHILDES database (see Table 2, above). The artificial language is outlined above. In these simulations the word-level categories contained 36 verbs, reflecting the number of verbs in Table 2, and 36 nouns. It was decided to keep the number of nouns equal to the number of verbs in order to avoid disparity between the code-length necessary for different sentence types. There were two sentence-types (*NV* and *VN*) reflecting the transitive and intransitive contexts of the verb constructions. Ten verbs were blocked with all 36 nouns for each sentence type (see Table 2), resulting in a total of 720 disallowed sentences.

Two of the four-element models described above were exposed to increasingly large corpora of this language. The first model contained word-level information about the 36 nouns and verbs, and sentence-level information about the NV and VN sentence types, but the exceptions element was empty: it did not contain any information about the 720 disallowed sentences. In this respect it was analogous to a learner who has acquired knowledge of word categories and sentence production rules, but has not learned that some sentences are illegal. This model would therefore be prone to overgeneralizations such as *I disappeared the rabbit.* The second model, by contrast, did contain information about the disallowed sentences. This model therefore required considerably more bits to specify initially, but the number of bits required to specify each sentence

of the corpus was fewer. In addition, a language learner who had learned these exceptions would not make the same overgeneralization errors that the first model would. Table 9 shows the relative simplicity of each model for increasingly large corpora as measured by the number of bits necessary to encode the model and the corpus data.

Corpus Size (sentences)	Model 1: Codelength (bytes)	Model 2: Codelength (bytes)
0	**0.1**	7.6
4000	**45.4**	51.1
8000	**90.8**	94.7
12000	**136.2**	138.3
16000	**181.5**	181.8
20000	226.9	**225.4**
24000	272.2	**268.9**

Table 9. Code-lengths of Models 1 and 2 for successively large corpora. Code-lengths in bold show the shorter codes for the corpus size.

It can be seen that for relatively small corpora (up to about 16,000 sentences), Model 1 gives a simpler encoding: less bits are required. For a learner who had heard relatively few alternation constructions, therefore, the tendency would be to code the data in these terms, resulting in overgeneralizations. For a more experienced learner, however, the simpler encoding would be that shown by Model 2, which requires fewer bits to encode relatively large corpora.

Conclusions and future directions

These results provide an initial confirmation that simplicity may provide a guiding principle by which some aspects of language may be learned from experience without recourse to a specific language-learning device. However, the simulations presented here are coarse-grained approximations of both the language and the language learner. Children do not process the language in batches of several thousand utterances. The models presented here were neither exposed nor sensitive to different word-type frequencies. A number of further studies which would provide considerably firmer support for the simplicity principle as a driving force for language acquisition suggest themselves.

Firstly, mathematical results show that word-type frequencies are important to the simplicity-driven learner, in that they may be the key as to when it becomes advantageous to posit exceptions to rules. Chater and Vitányi (2001) show that languages are approximately learnable given sufficiently large amounts of data. The CHILDES data in Table 2 therefore provides an indication of the order in which one would expect the learner to cease overgeneralizing words. An examination of children's speech that confirmed this order would be a major step towards providing robust support for the simplicity principle in language. Secondly, it would be useful to compare the timescale of recovery from overgeneralization in children with that of the model. This could be done by an examination of CHILDES database to determine an approximate relation between a child's age and the number of transitive/intransitive alternation constructions to which they have been exposed. It would then be possible to compare the learning rate of the child with that of the model. Again, this would be a useful source of evidence concerning the simplicity principle in language.

This chapter presented an alternative to Gold's idealization of the problem of language acquisition. It is suggested that there is sufficient statistical information in the input for a learner to learn quasi-regular alternating structures. These results are achieved by choosing the model of the language that provides the simplest (shortest) description of the linguistic data that has been encountered. These results re-open the question of the viability of language learning from positive evidence under less than ideal conditions, with limited computational resources and amounts of linguistic data available. They therefore also bear, indirectly, on the arguments concerning the balance between nativism and empiricism in language acquisition. More concretely, I suggest that the working hypothesis that the search for simplicity is a guiding principle in language acquisition deserves serious attention.

Chapter 8

Acquisition and Evolution of quasi-regular languages: Two puzzles for the price of one

The logical problem of language acquisition discussed in the previous chapter can be seen as the starting argument for raising a paradox about the evolution of natural languages: Firstly, if quasi-regular structures in languages are such hard cases for the learner, why are they so pervasive in contemporary natural languages? More specifically, why do not we see the emergence over time of simpler, more easily learnable languages? Secondly, the speculation that irregularities should tend to be replaced by regular forms over time leads immediately to a second puzzle: how did such language become quasi-regular in the first place?

This chapter falls into 3 main sections. Having discussed the ubiquity of quasi-regular constructions in the previous chapter I firstly discuss here the relationship between acquisition and evolution, in particular the idea that any hard learning problem of culturally transmitted information entails evolutionary puzzles. Secondly, I detail several simulations based on an Iterated Learning Model (ILM, e.g. Kirby, 2001) in which a probabilistically generated artificial language is transmitted over 1,000 generations of simplicity-based learners. The results of these simulations chart not only the stability but also the emergence of quasi-productivities in the language. In particular I show that:

a) Exceptions are stable across successive generations of simplicity driven learners.

b) Under certain conditions, statistical learning using simplicity can account for the emergence of quasi-productivity in a language.

In the final section I discuss the results of the ILM simulations, in particular the conditions in which quasi-regular structures might emerge.

The logical problem of language evolution

In this chapter I consider two questions for language evolution raised by the existence of idiosyncrasies. The first is a problem of transmission: what kind of learning mechanism could ensure the stability of idiosyncratic absences across generations and be sufficiently flexible and general to pre-date their emergence? The second is one of emergence: even assuming that such a mechanism exists, what conditions might give rise to these irregularities?

Simplicity-Based Language Learning: The Learner as Gambler

Chapter 7 showed that a batch learner – i.e., a learner that runs all calculations, after the entire corpus has been encountered – employing this strategy is able to distinguish genuine constructions from blocked ones as a result of exposure to data from the CHILDES database of child directed speech (MacWhinney, 2000). Here, I implement an online version that is able to postulate exceptions and create new hypotheses during the course of exposure to a rudimentary toy language. Algorithmic details are given in Appendix A; the following two sections describe the toy language and the learner's ability to discover exceptions in it.

The simplicity principle, outlined above, demonstrates how the simplest model of experience can be thought of as that represented by the shortest binary code. In this instance the binary code must represent two things: firstly a hypothesis, or grammar, that describes the language to which the learner is

exposed. Secondly, all the language that has been heard must be represented *under the hypothesis*. This may be expressed formally:

$$C = C(H) + C(D|H) \qquad [2]$$

Where C is the total length of code (in bits), C(H) is the number of bits necessary to specify the hypothesis (grammar) and C(D|H) is the number of bits necessary to specify the data (all the language heard) given the hypothesis. The length of code necessary to represent data will differ between hypotheses.

My model of the learner does not acquire vocabulary or induce categories and rules from scratch. I take productive rules to be already learned. Thus my model is already at the stage at which children make over-general errors. The task is to spot which of the constructions allowable under the rule are in fact blocked--- to find the holes in the language. Learning proceeds by a series of "gambles." The learner bets that a particular construction is not allowed and that it will therefore never be encountered. In making this gamble it must specify the construction as part of a new hypothesis, H. Coding this specification requires some bits of information, so the complexity of the new hypothesis increases. However, the learner has reduced the number of allowable constructions that it can expect to encounter. It has therefore increased the probability of those remaining. The number of bits required to specify future data under the new hypothesis is therefore reduced. Thus, if it is true that the construction is not allowed, the learner will gradually win back the number of bits that it gambled in specifying the exception. As more language is heard the new hypothesis will eventually

come to be associated with a shorter codelength than the original. If the gamble is inappropriate, however, the learner will encounter a construction that it has wrongly presumed to be disallowed. This is associated with a probability of 0, and hence an infinite code-length, so the 'gamble' is abandoned. my model generates a new hypothesis every time it gambles on a particular construction, with all hypotheses running in parallel. The preferred hypothesis is always that associated with the shortest codelength.

Learning a rudimentary language

A toy language was used to simplify the simulation. It was comprised of two syntactic categories, *A* and *B*, and two production rules, S_1 and S_2. The categories *A* and *B* each contained four words. The language also contained an exception element, specifying sentences that were producible under the re-write rules but were disallowed. Each sentence contained only two words, *AB* or *BA*. The language may be expressed formally as in [3]:

$S_1 \rightarrow AB$,

$S_2 \rightarrow BA$,

$A \rightarrow \{a_1, a_2, a_3, a_4\}$,

$B \rightarrow \{b_1, b2, b_3, b_4\}$,

$* \rightarrow \{(a_2),(a_2b_2),(a_2b_3),(a_2b_4),(b_1a_1),(b_2a_1),(b_3a_1),(b_4a_1)\}$ [3]

where the examples generated by * are blocked. This language can mimic the pattern of alternations, for example transitive and intransitive verb constructions.

In English, verbs can nominally occur in either a transitive or an intransitive context, but some are blocked from occurring in one or the other. This is analogous to the patterns in my toy language, where items in either category may in principle occur in either the first position, but can be blocked from doing so by entries in the exceptions element. This is illustrated in Figure 25.

Transitive	Intransitive
cut (Icut the cake)	*I cut
*I fell the bicycle	fall (*I fell*)
break (I broke the cup)	break (The window broke)

AB	BA
a_1 B	*Ba_1
*a_2B	Ba_2
a_3B, a_4B	Ba_3,Ba_4

Figure 25. The structure of the toy language mimics that of Baker's Paradox for alternations. a_1 and a_2 could be blocked from occurring in BA and AB constructions respectively by entries in the exceptions element such as a_2b_1, a_2b_2 or b_1a_1, b_2a_1 etc. For the first generation agent in each simulation, however, all As occurred in both contexts (that is, they were 'alternating'). 'Cut', 'fall', and 'break' are examples of alternating and non-alternating verbs. Levin (1993) provides an extensive list of alternations in English.

Samples of the language were produced by a parent agent and experienced by a learner agent. I assume that parents and learners share knowledge of word frequency. This allows both to associate each word with a probability of occurrence. Sentence probabilities are taken to be the product of two probabilities:

that of the first word and that of the second word, given the first. Parent agents use these probabilities to produce samples of the language stochastically. Learners use them to calculate codelengths (in bits) for different hypotheses. I assume that word frequencies are distributed according to Zipf's law (Zipf, 1948), an ubiquitous power law distribution in natural language (Bell, Cleary & Witten, 1990): If we rank words in terms of frequency, then frequency of any word is the inverse of its rank. Details are given in Appendix A.

Learner agents begin with a single, completely regular hypothesis about the language i.e., all sentences are allowed. This is equivalent to [3] with the exceptions element empty. As they experience samples of the language, the learner agents compare the probability of each sentence with the total number of sentences they have heard. A new hypothesis is generated if the total exceeds a threshold (where the threshold is a function of sentence probability; thus the threshold is different for each sentence). Each new hypothesis is simply a clone of the most recent hypothesis to be generated (or the original, if it is the first) with the addition of the sentence in question to the exceptions element. This addition entails an increment in the codelength associated with the new hypothesis, and a re-scaling of the probabilities for the remaining sentences.

Each sentence encountered entails an increment in the number of bits associated with each hypothesis, but since the creation of a new hypothesis involves rescaling sentence probabilities, this increment differs between hypotheses. All algorithmic details are given in Appendix A. Figure 26 illustrates the codelengths associated with all the hypotheses entertained by a learner agent after exposure to 50 sentences of a language containing 11 exceptions.

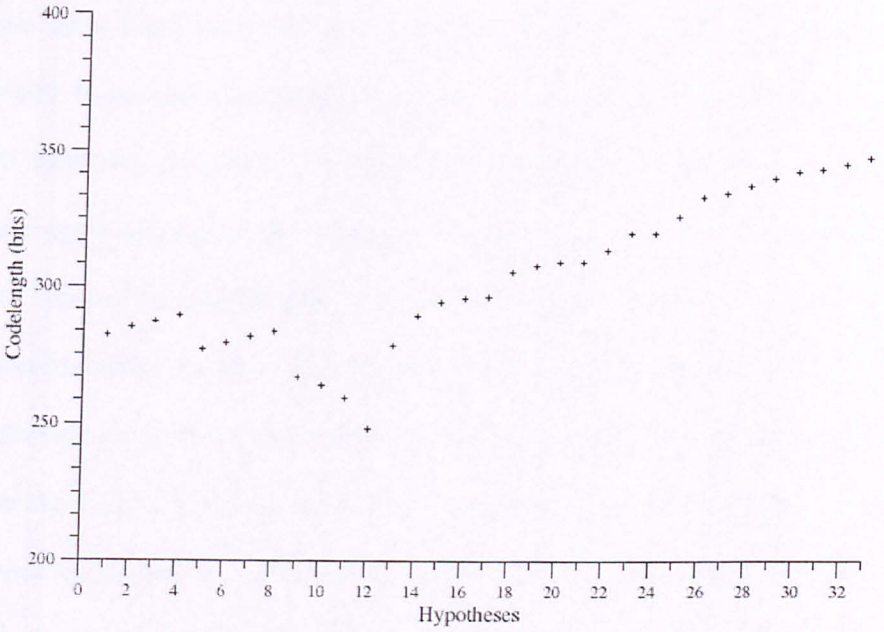

Figure 26. The codelength (number of bits) associated with each hypothesis grammar entertained by a learner after exposure to 50 sentences of a language containing 11 exceptions. The shortest codelength is obtained by the 12th hypothesis, i.e. the one containing 11 exceptions. (the first contains none, being completely regular). Although it is not obvious from the figure, the 12th hypothesis specifies exactly the 11 exceptions contained in the language.

Figure 26 illustrates that the learner agent creates many hypotheses, but that the shortest codelength is associated with the one that matches the language to which it was exposed. It is important to note that the sentence comprised of the two least frequent words was associated with a probability of approximately 1/100. It was therefore highly unlikely that it would have occurred in a corpus of 50 sentences. In addition, the learner received no feedback on its learning, other than more samples of the language. These conditions mirror to a modest extent the 'poverty of the stimulus', according to which children never hear all the possible sentences of a language and do not typically receive explicit negative feedback. In addition, our language contains, of course, no semantics and has no communicative function: I do not attempt to model the relationship between meanings-signals-referents nor try to give functional explanations of language change as in other models. In general, however, part of the fascination of the constructions investigated here is that their idiosyncrasy does not seem to be primarily semantically or functionally determined.

In spite of these restrictions, the learner agent was nonetheless able to distinguish between admissible and inadmissible sentences which it had not heard. It is also worth noting that this mechanism need not be restricted to spotting the idiosyncratic absence of single sentences: the same process could equally well be used to recover from overgeneral errors made as a consequence of (for example) semantic contexts. To see why this is so, it is helpful to consider how the sentences allowable under a grammar such as [3] can be represented in a contingency table:

	a_1	a_2	a_3	a_4	b_1	b_2	b_3	b_4
a_1	*	*	*	*				
a_2	*	*	*	*	*	*	*	*
a_3	*	*	*	*				
a_4	*	*	*	*				
b_1	*				*	*	*	*
b_2	*				*	*	*	*
b_3	*				*	*	*	*
b_4	*				*	*	*	*

Table 10. Sentences allowable under [3]. Rows are first words, columns are second words. The re-write rules license half the sentences in this table; blocked sentences are denoted *. The learner was able to discover exceptions to the rules such as a_2 appearing in first position or a_1 appearing in second position.

It is suggested here that a simplicity-based learning mechanism such as that outlined above is sufficiently powerful and general to offer a solution to the first of the evolutionary questions I posed, namely the transmission problem – i.e. once quasi-regularity is established, a learner can, in principle at least, learn this quasi-regularity, avoiding overgeneralization by using the simplicity principle. I now place the single learner in the context of an Iterated Learning Model (ILM) to consider the second question: conditions for the emergence of such idiosycrasies.

Language Learning over Generations - ILM simulations

Although developed independently, my model proposes an MDL learner embedded within an Iterated Learning Model (ILM), which has been used extensively by Kirby and colleagues, and others (e.g. Kirby, 2001, Brighton, 2002; Teal & Taylor, 2000; Zuidema, 2003). In the ILM, parent agents generate language for children agents, who in turn, become parents for the next generation

of learners. A simplifying assumption is that there is one agent per generation, so issues of population dynamics are neglected. All agents were "genetically" homogeneous, i.e. all were equipped with identical learning facility and started from the same point in their development. The first generation agent was exposed to probabilistically generated samples of the completely regular toy language used in the single-learner simulation. Subsequent agents were exposed to probabilistically generated samples of the language as learned by the preceding generation. Although complete regularity at the outset is probably unrealistic, my intent is not so much to replicate an historic development of languages as to test the conditions for the emergence and stability of irregularities. I test this in the least favourable condition for their emergence, i.e. an ideal fully regular language. The mean number of sentences heard by each agent was the same within each simulation, but varied between simulations. In different simulations, successive generations of learners heard between 25 and 65 sentences. Again, it was unlikely that any agent was exposed to all the sentences in the language, and agents received no negative feedback on their learning. When an agent had been exposed to the required number of sentences, one hypothesis entertained by that agent was selected. This hypothesis was then used as the basis for generating the sentences that would be heard by the succeeding generation. The hypothesis chosen was always that associated with the simplest interpretation, i.e., that with the shortest code length.

Results

Figure 27 charts the emergence and stability of exceptions in four simulations. The number of sentences heard by each generation was critical to both. Where each generation heard a short corpus (mean number of sentences, n, of 30, Fig. 27(a)), exceptions frequently emerged but were highly unstable: they rarely remained in the language for more than a few generations. With a long corpus (mean n=60, Fig. 27(d)) exceptions were less likely to emerge; in contrast to Figs 27(a) - 27(c) no exceptions emerged for almost 400 generations. However, once they had emerged they were much less likely to be lost from the language than with shorter corpora.

Figure 27 suggests that exceptions are posited during the early stages of language acquisition. With a relatively small amount of data, learners may postulate that the language contains many exceptions that do not in fact exist. As more data becomes available, such early hypotheses are either confirmed or exposed as spurious. These simulations suggest a trade off between emergence and stability of exceptions. The crucial factor mediating this trade off is the amount of language heard by each generation. If each generation hears a great deal of data, exceptions are unlikely to emerge: any that are posited will later be shown to be false. However, if exceptions are to be stable, each generation must hear enough language to learn the exceptions that existed in the previous generation.

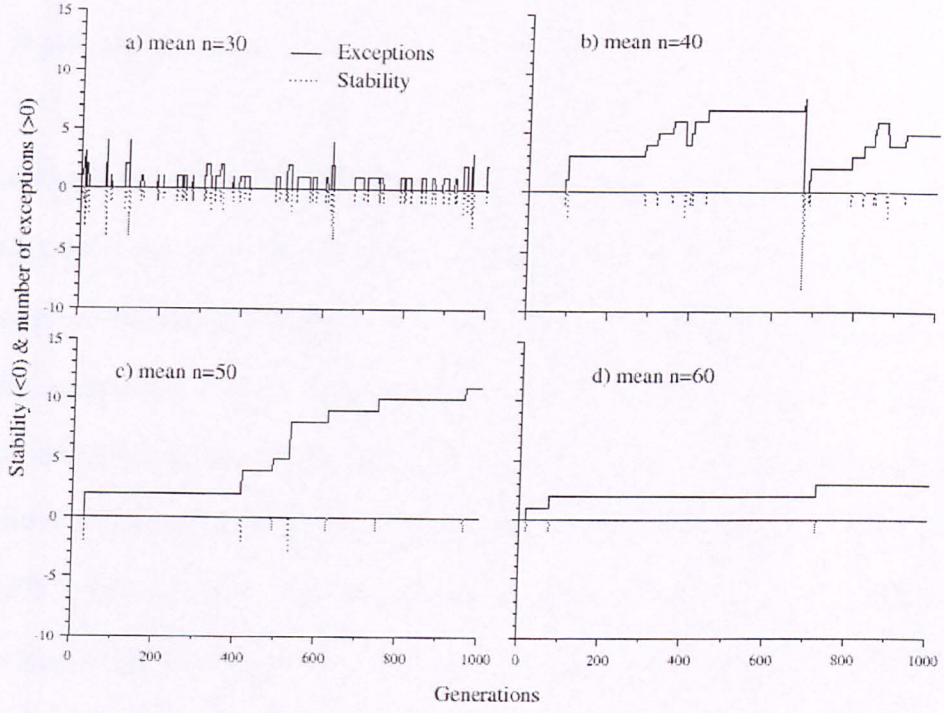

Figure 27. The number of exceptions contained in the language transmitted by each learner to the subsequent generation in four simulations with differing corpus sizes. Where the number of exceptions was stable across several generations, for example seven exceptions in c) or the final 600 generations of d), the sentences specified as exceptions were the same for each generation. It is important to note the difference in scale for number of exceptions for a), b), c) and d).

Discussion and conclusion

In this chapter and in Chapter 7 I started by noting that most phenomena in natural languages seem to be of a quasi-regular nature, which traditionally poses a learnability problem. Baker's paradox arises whenever the child has to recover from perfectly plausible and attested overgeneralisations such as (Fisher, 1976):

**I gave my mummy it*

without the aid of direct negative evidence. Because a putative Universal Grammar can only capture general syntactic behaviours, it looks like most syntactic constructions have to be learned from experience. I contended that if the acquisition of such idiosyncrasies is hard, then their transmission over generations of speakers should be 'filtered out' over time to improve learnability and communication. I subsequently presented a computational simulation where such hard cases are in fact successfully learned and transmitted from positive evidence. My solution to the learning problem is that a learning bias toward simplicity of representation makes language learnable from experience. This bias need not be specific to language –indeed simplicity principles have been used in the context of linguistic (Brent & Cartwright, 1997; Goldsmith, 2001 Wolff, 1982) and non-linguistic contexts (e.g., perception, Hochberg & McAlister, 1953; van der Helm & Leeuwenberg, 1996; categorization, Pothos & Chater, 2002), and have even been viewed as general frameworks for cognition (e.g., Chater, 1999; Wolff, 1991). In my model there is no *a priori* 'correct' grammar, i.e. a grammar that is valid prior to linguistic experience. The development of the

final-state grammar corresponding to adult linguistic competence is a matter of choosing the simplest competing grammar.

The quest for simplicity is hardly a new idea and appears, for instance, in the early works of Chomsky (1955; 1965: 25): under the notion of markedness the grammar being constructed directly reflects the linguistic input. If the input contains information that points to a certain complex grammatical relation, the learner will acquire it, but if the input lacks such information, the principles that govern generalization will prevent the learner from constructing the more complex grammar. The markedness approach was abandoned in generative linguistics, in part because of the lack of a metric for establishing the simplicity of grammars, and partially for the rise of the 'poverty of the stimulus argument' whereby linguistic experience seems hopelessly unreliable. Such caveats are dealt with in this chapter: firstly, the MDL approach provides a quantitative metric for simplicity; secondly, the poverty of the stimulus instantiated in the transmission bottleneck seems a necessary precondition for the emergence of exceptions rather than a hindrance to language evolution. There is a critical size for the bottleneck: too little or too much exposure to the language fails to yield stable patterns of quasi-productivity[13].

Another defining feature of the simulations described in this chapter is that they rely on word frequencies to assign probabilities to sentences. I have also assumed that the distribution of word frequencies follows Zipf's law (Zipf, 1948). These assumptions merit some discussion. There are two important reasons for applying a power law distribution to word frequency: firstly, it has

[13] Brighton (2002) and Kirby (2001) found that both compositionality and irregularity emerge thanks to the bottleneck. Interestingly, I seem to have modelled the reverse timecourse of Brighton's simulations, which start with a non-compositional language to attain compositionality. The converging end-point is, however, a stable state of quasi-regularity modulated by the bottleneck.

been shown in the past to have important implications for the emergence of irregularities in ILM simulations of language evolution (e.g. Kirby, 2001), and secondly, such frequency distributions are ubiquitous in natural language.

Kirby (2001) has shown that benign irregularities[14] will spontaneously emerge in compositional language structure if frequency distributions follow Zipf's law. When this is the case, the very frequent phrases at the 'head' of the distribution are shortened to irregular forms, resulting in selection under a similar MDL metric as that described here. This phenomenon does not appear to occur when frequencies do not follow a power law. We can see the impact of Zipf's law on the simulations by considering the likely results if word frequencies had been evenly distributed (i.e. if all sentences had been assigned equal probability). In such a case, the threshold number of sentences for learning a particular exception would have been the same for every sentence. Thus the learner would either encounter enough sentences to learn all the exceptions at once, or would not learn any exceptions at all. Any sentences not encountered before the threshold was reached would be posited as exceptions. It is not impossible that exceptions would emerge and survive under such conditions, but it seems unlikely that we would see the patterns of emergence and stability outlined above.

In following Zipf's law, the frequency distribution of words in my toy language mirrors that found in natural languages: word frequencies in natural language text and corpora follow such distributions quite precisely, as do a

[14] whereas I investigate the case of accidentally unfilled slots in syntactic paradigms, Kirby models the case of slots filled by irregular forms, e.g. the emergence and replacement of *went* for **goed*. Baker called these 'benign' exceptions vis-a-vis the learnability paradox: recovery from overgeneralisation of **goed* can be safely arrived at by positive evidence, as the correct alternative *went* is present in the input. In addition, Kirby models meaning, and the pressure to invent random forms for meanings for which no rule exists is what gives rise to the irregularities in the first place. Because I purposely modeled the emergence of quasi-productivities without a meaning space, comparisons with Kirby's work can only be indirect.

number of other natural language statistics (Bell *et al.*, 1990). The assumption that the probability of a given sentence is perceived as a function of word frequencies is more controversial. It seems highly unlikely that this would be exclusively the case in natural language; I would be surprised if factors such as semantics and phonology did not play a role. However, no factors other than the frequency and collation statistics were available in the language. I contend that it is a plausible assumption that these factors also play a role in determining our perceptions of the probability of a particular sentence occurring. I speculate that in the absence of other factors they must determine them exclusively.

Anecdotally, it seems that young speakers are losing the Germanic/Latinate distinction that allows Dative shift for *give* but not for *donate*. Hence **John donated the library a book* is more likely to be accepted as grammatical in contemporary usage. However, **John said Mary hello* is more recalcitrant to regularization, perhaps because *donate* is a low frequency verb whereas *give* has a high frequency. In the group of collaborators I work with we have ourselves found that our intuitions concerning 'holes' in the language are surprisingly volatile – we find it hard to reject some of the ungrammatical examples we have used several times as examples in our discussions. The same 'lifelong learning' phenomenon also affects linguists who feel that subjacency violations become weaker the more often they produce them (Culicover, 1995). This is consistent with my model. In addition syntactic constructions such as Dative shift may undergo local regularization while still preserving idiosyncratic behaviour in some other area (*waved/say hallo*, or *send/report*). More interestingly, my simulation results defy intuition in that a reverse trend from local regularity to idiosyncratic behaviour can also occur.

Although relatively stable, a given idiosyncrasy may die out quickly leaving the place to new ones or to a regularized form. Local structural reorganizations of syntactic paradigms (such as Dative shift for *donate*) can take place within a *single* generation. An implication of my model, not tested directly, is that linguistic diversity will emerge spontaneously in different spatially distributed linguistic communities, even in those that share a similar culture, as attested in different varieties of English in the English-speaking world. These considerations remain speculative as I have not attempted to model language change driven by social factors, language contact, multilingualism, or other factors.

In this chapter I have shown that a potentially hard problem of language acquisition, that of quasi-regularity, gives rise to a paradox of language evolution. The acquisition problem may be solved by incorporating a learning bias towards simplicity. This solution goes some way towards resolving a related paradox in language evolution: given sufficient exposure to samples of language, quasi-regular structures are learnable, and hence stable over generations. In addition to this I have shown that under some conditions, quasi-regular structures may emerge in a language even if it were initially completely productive. However, I make no assumptions as to the origins of language in the human species. The starting point of a fully regular language should not be taken as an hypothesis about historical languages. It rather served the purpose of demonstrating that quasi-regular structures may emerge spontaneously, and hence constitute a natural stable equilibrium for languages across time.

It is worth mentioning the striking analogy between natural languages and many complex systems in the natural world. The sciences of complexity have

recently started to note that most natural phenomena are truly complex, i.e. they occur at a transition point between two extremes, perfect regularity on the one side and pure randomness on the other (Flake, 2001). Perfect regularity is orderly and allows for high compressibility, whereas strictly irregular things are random and cannot be compressed because completely unpredictable (Gell-Mann, 1995). If syntactic constructions were completely idiosyncratic (irregular) they could only be learned by heart and no generalisation to novel instances would be possible. On the other side, the sort of innate constraints for acquisition postulated by a Universal Grammar and characterized in terms of maximally general and universal syntactic principles would lead all languages to develop perfectly compressible grammars, which is not the case for natural languages in the world. For example, a truly general transformational rule like Dative shift movement raises the projection problem noted by Baker, as it predicts that **We reported the police the accident* is grammatical. Hence, it is ultimately contended that the very nature of irregular, idiosyncratic, and quasi-regular forms so widely spread and stable in natural languages suggests that they are arbitrary and unconstrained except by the requirement that they be computable, i.e. learnable (see also Culicover, 1999). A language learning mechanism must be capable of accommodating the irregular, the exceptional, and the idiosyncratic. I have proposed that a general-purpose learning mechanism driven by simplicity has the computational power to do so.

Chapter 9

Discussion and conclusions

In this thesis I have attempted to extend empirically and computationally the basis for a reappraisal of probabilistic accounts of language learning. Traditionally, statistical learning in the classical associative sense has been associated with behaviourism and hence been downplayed as having quite limited power. Knowledge built out of relations based on temporal and spatial contiguity can account for pattern recognition and memory retrieval based on similarity assessment, but not, it seemed, for abstract structural dependencies such as nonadjacencies and phrase structure. The naïve view of statistical learning has pushed researchers in search of more sophisticated computational tools (Fodor & Pylyshyn, 1988). In language in particular, the domination of the rationalist position with Chomsky has caused serious resistance to accept statistical learning as a viable research project. However, much of the work in this thesis contributes to the idea that statistical learning need not be as naïve as it is portrayed by its detractors.

The agenda for statistical learning can be divided into two main concurrent lines of enquiry: the first investigates the probabilistic nature of the input and the availability of reliable statistical cues potentially available to the learner. The contribution of the present work provides sound evidence for a cascade of potentially useful statistical cues – several yet to be discovered – that span from n-gram statistics to nonadjacencies, to perceptually salient acoustic and phonological cues. In this sense we can begin formulating an argument for the "richness of the stimulus", contrary to the received wisdom of the poverty of the linguistic stimulus. The second line of enquiry is a direct consequence of such richness in statistical cues. Given the combinatorial explosion of analysing

each possible combination of statistical relations in the input, learning must be guided by some powerful inductive principle imposing constraints on the possible interpretations (i.e. hypotheses, or grammars) and arbitrary dependencies from a finite amount of data. Positing statistical learning mechanisms does not necessarily imply commitment to a *tabula rasa* position. It is not psychologically realistic to assume that the learner will blindly search for all possible relationships between a vast range of properties. This forms a valid response to Pinker (1987) who, in criticising distributional methods, pointed out that a distributional analysis of sentences 1-3 below would lead the learner to incorrectly categorise *fish* and *rabbits* together and to overgeneralise to 4, and that these errors are not found in childrens' spontaneous productions.

(1) John ate fish

(2) John ate rabbits

(3) John can fish

*(4) *John can rabbits*

distributional analyses need not be as simplistic as those suggested in Pinker (1984). Redington, Chater, & Finch (1998) have convincingly argued that the fact that naïve "spurious correlations" based on single examples lead to erroneous generalisations does not rule out the entire class of distributional analyses, which is in fact more powerful.

The reduction of uncertainty hypothesis, the connectionist models implemented, and the simplicity principle advocated in this thesis can all be regarded as equivalent formulations – at some general level – of an inductive

principle that specifies what needs to be done, namely filtering the information available in search for a reliable and economical compression of the data. A comparison between such formulations is beyond the scope of this thesis. Neural networks and simplicity are formalised and computationally implemented specifications of a learning algorithm, while reduction of uncertainty can be regarded as a general working framework, underspecified computationally. On the one side, there is great overlap between such theoretical proposals: for instance, to the extent that the connectionist simulations can replicate the Variability effect they can be regarded as an instantiation of reduction of uncertainty. On the other side, the search for reduction of uncertainty as discussed in the early chapters focuses more on filtering among potential candidate sources (bi-grams, trigrams, nonadjacencies), whereas the simplicity principle was used in the simulations in chapters 7 and 8 to select among competing hypothesis grammars given a set of chosen linguistic elements. These differences, however, may only depend on which angle the researcher chooses to tackle the issue of learning from positive data.

Another issue that was tackled in the thesis is the orthogonality argument with respect to the traditional argument for the separation of statistical and algebraic styles of computation. Although statistical learning as a field of research tends to "flirt" more with associative styles of computation, it need not take a conclusive stance on the issue, at least not until experimental segregation can be shown to be effective (and the null results in chapter 6 would suggest caution against hasty interpretations). In fact, whereas I use a neural network to model the Variability effect in chapter 3, recovering from overgeneralisations

using simplicity specifies rules for verb category assignment and rules for exceptions and is implemented in standard symbolic programming.

Limits and future directions

Before concluding, I would like to highlight below some limits of the current work and suggest potential avenues for research to follow up.

Extensions to the variability effect

Here I provide a number of possible extensions to test the robustness of the variability effect. The first obvious next step would be to test the zero-variability condition on children, given that Rebecca Gómez found the variability effect on both adults and children in her original paper. At the time of writing Gómez is currently testing this hypothesis (personal communication).

Another experiment, the "infinite-variability condition" would test whether there is an optimal degree of variability (accidentally 24 embeddings), or whether the more variability the merrier. In this condition, each training string would appear with a new (unseen) embedding. If the hypothesis that learners disregard the embedding as irrelevant with large variability is truly correct, then having potentially infinite variability should not constitute an hindrance to both detecting nonadjacencies and generalising to new embeddings. If anything, it should make the task easier and performance could be better than in the Set size 24 condition.

Another control experiment would involve testing participants on six frames and one embedding. In this way the number of *type* strings would be the same as in Set size 2. If participants performed well, then it could be argued that

learners did not learn well in the zero-variability condition because they only had 3 type strings to memorise. Partly this control becomes less necessary given successful generalisation to novel middle items even in the Set Size 1 condition, as evidenced in Experiment 5, which suggests that the strings are not merely learnt by rote.

Yet another interesting new condition to test, the "asymmetrical grammar" condition, came up as a result of a conversation with Axel Cleeremans. Servan-Schreiber, Cleeremans, & McClelland (1991) addressed a similar problem of learning nonadjacent dependencies. Their connectionist simulations suggest that learners may be able to learn with a low variability of embeddings, say only 3 embeddings, provided that these have slighly different frequencies, e.g. X_1=40%, X_2=30%, X_3=30%. An analogous version would involve embeddings appearing with different frequencies with different nonadjacent dependencies. For example, the X_1 in the $A_1\_B_1$ frame would occur with a frequency of 0.4 and the same X_1 in the $A_2\_B_2$ frame would occur with a frequency of 0.6. Servan-Schreiber *et al.* argued that the rationale for these asymmetries in connectionist terms is that the recurrent networks preserve nonadjacent information better if the embedded material is statistically differentiated during training.

In general, two critiques could be levelled at the artificial language used here for detecting nonadjacencies. Firstly, although non-adjacent, the dependencies are still somewhat local because they only span one intervening word. Secondly, they always occurred in third and last position. Instead, nonadjacencies in natural languages can span several embedded words and typically occur in different relative positions. This non-fixedness of constituents'

relative position means that learners must ultimately abstract beyond the ordering of specific words. Because distributional methods have been criticised for being unable to accommodate free-ranging relative position, further experiments could test whether participants are able to detect nonadjacencies when the training items include strings such as *A x y z B* and *A x B y z*, with dependencies placed in different relative positions during training.

Regarding the connectionist simulations, these are far from providing a comprehensive and accurate picture of the variability effect. Firstly, the U-shaped results from the generalisation experiment could not be simulated entirely – performance was low on Set size 2 and high on Set size 1 as expected, but not high on Set size 24. Secondly, the simulations did not capture the differences in performance found across modalities. Both problems might be overcome by changing the input and output units from localist to semi-localist or distributed. Localist encoding is not the best way to elicit correct generalisation in neural networks. From the point of view of the network, the new middle item is a completely new vector that bears no resemblance whatsoever with previous vectors. This is equivalent, in human experimental settings, to showing a completely unrelated item as new embedding at test, say the picture of a cow. Now it is reasonable to assume that human participants would have a hard time deciding whether a pseudo-sentence *pel_<picture of a cow>_rud* was grammatical, independent of the correctness of the frame. Hence, distributed representations may be a better way of encoding key features of the stimuli common to all other stimuli, for instance phonological properties. In the same vein, performance discrepancies found across modalities could be found by coding the difference associated with hearing a stimulus as opposed to seeing it

on the screen. Overall, it is not known at present whether connectionist models as a general class of learning models could scale up to the complexity of real natural languages. And it is not immediately obvious that simple recurrent networks can represent nonadjacent dependencies that are free to range in relative position, as opposed to coding a specific position, as it was the case with the A-X-B language used in my simulations. In most natural language sentences, the material that separates two nonadjacent constraints is hardly of the same length, so a harder testbed for SRNs would be to learn such cases.

What is learnt in Artificial Grammars

In devising an AGL experiment using finite-stage grammars the experimenter decides what is the correct set of responses, although several grammars might generate the set of data that participants are trained on. Take as an example the *A_X_B* language used in the variability experiments. Participants were asked to discriminate between $a_1\_x_2\_b_1$, and $*a_1\_x_2\_b_3$, although *positionally* the ungrammatical string does not violate any rule. One could have easily conceived of a set of rewrite rules that construct the *very same* training stimuli by simply indicating the position of items in sentences. Specifically:

$S \rightarrow A\ X\ B$

$A \rightarrow a_1, a_2, a_3$

$X \rightarrow x_1, x_2$

$B \rightarrow b_1, b_2, b_3$

It is possible to further assign equal probabilities to each element:

$A \rightarrow a_1\ (.33),\ a_2\ (.33),\ a_3\ (.33)$

and to subsequently choose a subset of all possible sentences generated by the rewrite rules such that only *As* and *Bs* with same subscript (e.g. $a_1\_x_2\_b_1$, but not $a_1\_x_2\_b_3$) are included in the training set. Notice that this new training set would be exactly the same as the original one used by Gómez and in this thesis. At test, the experimenter might be interested in testing whether participants have generalised positional information instead of nonadjacencies, In this case the forced choice test might require a distinction between $a_1\_x_2\_b_3$ versus $*x_2\_\ a_1\_b_3$. Notice that the string that in my experiments was considered ungrammatical now has become the grammatical one. However trivial this example might seem, it suggests that the notion of what is grammatical and ungrammatical in AGL experiments and the notion of underlying rule dictating allowable sentences may be sensitive to what choices participant are required to make their judgements upon. In the eye of the necessarily naïve participant the same training items may be classified under differentially interpretable but overlapping patterns. In particular, it is possible that participants implicitly and concurrently entertain different hypotheses, which they disambiguate at test on being prompted with the forced choice task. Anyhow, it would be a mistake to assume that participants who scored poorly on the variability task are not able to detect any sort of structure. A more plausible explanation is that they may have picked up on different patterns, for instance the one specifying positional information outlined above. For these participants, $a_1\_x_2\_b_1$, and $a_1\_x_2\_b_3$ are both positionally correct, which would explain the higher than chance ratio of yes responses in low

variability conditions. After-test verbal reports carried out informally during my experiments suggest that some learners were focusing on positional information. For instance, in Experiment 1, one participant said "The rule is that *wadim* always occurs in middle position". These considerations do not undermine the results of course, but invite us to consider what is learnt in AGL more carefully. Specifically, theoretical questions such as "are there rules or statistics in language learning?" or "can participants learn this structure?" progressively loose interest in favour of questions such as "given several potential cues and interpretations available, under what specific conditions do learners converge towards the same structure hypothesis?" Absolute positional information is not less worthy a hypothesis about the potential organisation of a set of stimuli than other forms of structure, for example nonadjacent structure. However, some types of structure may be more affordable or perceptually more salient than others: for instance, learners may be naturally biased toward classifying stimuli based on positional information as a *default* hypothesis. We have seen that this may be particularly true of artificial grammars containing two-word long sentences (e.g. Smith, 1966). Smith found that learners acquire only absolute positional information in his AGL study that explored the role of lexical co-occurrence patterns as a means for abstracting category structure without form-based cues. In addition to these considerations, learners may be more prone to pay attention to bi-gram transitional probabilities, and then trigrams, before focusing on nonadjacencies. As an example, in English whether a noun precedes or follows a verb almost invariably determines whether it is a subject or an object. Consider *John ate the broccoli.* It is true that to abstract categories such as NOUN and VERB learners must acquire representations that are independent

of context, for instance in hearing the word "joy" outside context one immediately knows it is a NOUN. But because nouns and verbs typically appear in patterned contexts, positional information – which is most readily available to learners – may be the cue they pick up on in the first place and may represent a first step towards building more abstract categories (see Gómez & La Kusta, in press). Besides, positional information in English is an essential component to distinguishing meaning, for instance agent and patient thematic roles in sentences like *John ate the broccoli* and *The broccoli ate John.*

Following from the above considerations, a project deserving more research is thus whether hypothesis testing in AGL and natural languages is modulated by *hierarchies of cues*, in the sense that learners might prefer some hypotheses over others as default. Also, do multiple cues augment learners' ability to detect structure or do cues act in a winner-takes-it-all manner where one cue prevails over the others? For instance, in the segmentation experiments of chapter 6 learners' preference for plosive sounds in word-initial position seemed to annihilate the contribution of distributional structure. Likewise, in the generalisation experiments of chapter 6 the plosive sound cue seems to be suppressed by a preference for syllables following silent gaps, however small the gaps may be and whatever phoneme follows the gap. That distributional cues may be overridden by perceptual cues such as stress patterns and coarticulatory cues has also been documented in Johnson & Jusczyk (2001).

In addition, what happens in the presence of conflicting cues? How can learners distinguish relevant from irrelevant linguistic input? Gómez and Lakusta (submitted) found that learners are capable of suppressing noise in the input up to a limit. Although all these questions will not find an answer in this work, my

results seem to point towards an increasing role of the environment in guiding learning and to the fact that learning from the environment is much more powerful than previously acknowledged. In particular, the connectionist simulations suggest that simple associative mechanisms may be powerful enough to detect nonadjacent dependencies.

Another relevant issue to take into consideration is how experimental instructions interact with training and test procedures and may indirectly guide learners' choices at test. For instance, participants' knowledge of the subjacent structure may be very fuzzy up to the point of test instructions telling them that half of the strings they are going to hear are ungrammatical. At test, participants focusing on positional structure may find all sentences grammatical because they all conform to the rule. Some of these may switch from positional to nonadjacent structure after inferring that the first few test items encountered "can't just be all grammatical". In this case, perhaps a Signal Detection Analysis would reveal a high ratio of inconsistencies between first and second trial, with a high Error/Correct ratio. The case for "learning" taking place at test has been made by Redington & Chater (2002). Although definitely speculative, these considerations are a first step towards a more ecologically aware methodology of experimental testing.

Solving the language acquisition and evolution puzzles with Artificial Grammars

The simulations presented in chapters 7 could find a natural follow-up in a series of AGL experiments that would test whether it is possible to learn to generalise and, at the same time, constrain overgeneralisations. Using the now familiar *A_X_B* language, one could think of three frames, arbitrarily corresponding to

three abstract syntactic patterns. The embeddings would represent the class of possible words (e.g. verbs) allowed under each pattern. Some embeddings could be matched with all three patterns while some others would either be associated to only one or two of them. At training, participants would be showed a subset of all possible sentences with an associated probability distribution, for instance a Zipfian distribution. A forced choice recognition test would probe whether participants could generalise to unseen sentences, one with high probability of occurrence, the other with low probability. Using a between-subject design one condition might include a relative short training, in which one would hope to elicit overgeneralisations. The other condition would include a prolonged training using the same stimuli in the hope that indirect negative evidence in the form of a highly expected but never encountered string would have accumulated. The prediction is that learners would be able to judge successfully which one among pairs of possible strings would be more likely to be part of the language.

If this experiment turned out to be effective, an evolutionary experiment could be set up to replicate the findings of chapter 8, where each generation would be represented by a participant learning the AGL and subsequently transferring to a new learner. The participant would be asked to participate in an experiment of language survival, by listening to sentences from a newly discovered African language on the verge of extinction. They would be required to try to learn (or memorise) the language in order to transmit it to a new learner. Overcoming some (non-trivial) caveats associated with eliciting a set of sentences from each participant, it could be possible to feed each new generation/participant from the language produced by the previous participant. In

this way, one would hope to see the emergence and stability of syntactic holes in a similar way to the trend obtained in Figure 27 with the computer simulations.

References

Anderson, J. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum Associates.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1996). Computation of conditional probability statistics by 8-month old infants. *Psychological Science*, 9, 321-324.

Baayen, R.H., Piepenbrock, R. Gulikers, L. (1995). *The CELEX Lexical Database* (CD-ROM). University of Pennsylvania, Philadelphia, PA.

Baker, C.L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, 10, 533-581.

Baker, C.L. and McCarthy, J.J., eds (1981) *The logical problem of language acquisition*. Cambridge, Mass.: MIT Press

Bell, T.C., Cleary, J.G. & Witten, I.H. (1990) *Text Compression*. Upper Saddle River, NJ: Prentice-Hall

Black, A.W., Taylor, P., & Caley, R. (1990) *The Festival Speech Synthesis System*, available from http://www.cstr.ed.ac.uk/projects/festival.html, Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, UK.

Bod, R., Hay J., & Jannedy, S. (Eds.) (2003). *Probabilistic Linguistics*. The MIT Press.

Bolinger, D. (1968). *Aspects of Language*. New York, Harcourt, Brace.

Bowerman, M. (1982). Evaluating competing linguistic models with language acquisition data: Implications of developmental errors with causative verbs. *Quaderni di semantica*, 3, 5-66.

Bowerman, M. (1996). Argument structure and learnability: Is a solution in sight? *Proceedings of the Berkeley Linguistics Society*, 22, 454-468.

Braine, M. (1987). What is learned in acquiring word-classes--a step toward an acquisition theory. In B. MacWhinney (Ed.) *Mechanisms of language*, pp. 65-87.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life*, 8(1): 25-54.

Broeder, P., & Murre, J. (2000). *Models of language acquisition: Inductive and deductive approaches*. Oxford: Oxford University Press.

Brooks, P. & Tomasello, M. (1999). How young children constrain their argument structure constructions. *Language*, 75, 720-738.

Brooks, P. & Tomasello, M. (1999). Young children learn to produce passives with nonce verbs. *Developmental Psychology*, 35, 29-44.

Cartwright, T.A., and M.R. Brent (1997). Syntactic categorization in early language acquisition: formalizing the role of distributional analysis, *Cognition*, 63, 121-170.

Cassidy, K.W. & Kelly, M.H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language*, 30, 348-369.

Cassidy, K.W. & Kelly, M.H. (2001). Children's use of phonology to infer grammatical class in vocabulary learning. *Psychonomic Bulletin and Review*, 8, 519-523.

Chater, N. & Vitányi, P. (2001). A simplicity principle for language learning: re-evaluating what can be learned from positive evidence. *Manuscript submitted for publication.*

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, 103, 566-581.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *Quarterly Journal of Experimental Psychology*, 52A, 273-302.

Childers, J. & Tomasello, M. (2001). The role of pronouns in young children's acquisition of the English transitive construction. *Developmental Psychology*, 37, 739-748.

Brooks, P., Tomasello, M., Lewis, L., & Dodson, K. (1999). Children's overgeneralization of fixed transitivity verbs: The entrenchment hypothesis. *Child Development*, 70, 1325-37.

Chomsky, N. (1957). *Syntactic Structures.* The Hague: Mouton & Co.

Chomsky, N. (1995). *The minimalist program.* MIT Press.

Chomsky, N. (1955). *The Logical Structure of Linguistic Theory.* Manuscript, Harvard University. Published by Plenum Press, New York and London, 1973.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Chomsky, N. (1980). *Rules and representations.* Cambridge, MA: MIT Press.

Christiansen, M.H., & Chater, N. (1994). Generalization and connectionist language learning. *Mind and Language*, 9, 273-287.

Christiansen, M.H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23, 157-205.

Christiansen, M.H., Conway, C.M., & Curtin, S. (2000). A Connectionist Single-Mechanism Account of Rule-Like Behavior in Infancy. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, Mahwah, NJ: Lawrence Erlbaum Associates.

Cleeremans, A. Servan-Schreiber, D., & McClelland, J.L. (1989). Finite state automata and simple recurrent networks. *Neural Computation,* 1, 372-381.

Content, A., Mousty, P. Radeau, M. (1990). Brulex. Une base de données lexicales informatisée pour le français écrit et parlé. *L'Année Psychologique*, 90, 551-566.

Conway, C.M., & Christiansen, M.H. (2002a). Sequential Learning through Touch, Vision and Audition. Paper presented at the *24th Annual Conference of the Cognitive Science Society*, Fairfax, VA.

Conway, C.M., & Christiansen, M.H. (2002b). Modality Constrained Statistical Learning of Spatial, Spatiotemporal, and Temporal Input. Poster to be presented at the *43rd Annual Meeting of the Psychonomic Society*, Kansas City, KS.

Culicover, P. (2000). *Syntactic nuts.* Oxford: Oxford University Press.

Culicover, P. W. (1995). Adaptive learning and concrete minimalism. *Proceedings of GALA 95.*

Daugherty, K., & Seidenberg, M. (1992). Rules or connections? the past tense revisited. *Annual Conference of the Cognitive Science Society*, 14, 259--264.

Dienes, Z. (1992). Connectionist and memory-array models of artificial grammar learning. *Cognitive Science*, 23, 53-82.

Dowman, M. (2000) Addressing the Learnability of Verb Subcategorizations with Bayesian Inference. In Gleitman, L. R. & Joshi, A. K. (Eds.) *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society.* Mahwah, New Jersey, USA: Lawrence Erlbaum Associates.

Dulany, D.E., Carlson, R.A., & Dewey, G.I. (1984). A case of syntactical learning and judgement: How conscious and how abstract? *Journal of Experimental Psychology: General*, 113, 541-555.

Elman (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48, 71-99.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, 14, 179-211.

Fiser, J., & Aslin, R.N. (2002). Statistical learning of higher-order temporal structure from visual shape-sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 130, 658-680.

Flake, G.W. (1998). The computational beauty of nature. Cambridge, MA: MIT Press.

Frigo, L., & McDonald, K.L. (1998). Properties of phonological markers that affect the acquistition of gender-like subclasses. *Journal of Memory and Language*, 39, 218-245.

Gell-Mann, M. (1995*). The quark and the jaguar: Adventures in the simple and the complex.* New York: W.H. Freeman.

Gibson, E.J. (1991). *An Odyssey in Learning and Perception.* Cambridge, MA: MIT Press.

Gleitman, L. R., Gleitman, H., Landau, B. & Wanner, E. (1988). Where learning begins: Initial representations for language learning. In F.J. Newmeyer (Ed.), *Linguistics: The Cambridge survey*, Vol. 3, pp. 150--193. Cambridge, England: Cambridge University Press.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 16, 447-474.

Goldberg, A. (1999). "The emergence of the semantics of argument structure constructions". In MacWhinney, B. (Ed.) *The emergence of language.*

Goldberg, A. (2003). Constructions: a new theoretical approach to language. *Trends in cognitive sciences*, 7, 219-224.

Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2): 153-198.

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, 13, 431-436.

Gómez, R.L., & Gerken, L.A. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109-135.

Gómez, R.L., & Gerken, L.A. (2000). Infant artificial language learning and language acquisition. *Cognition, 70*, 109-135.

Gómez, R.L., and La Kusta, L. (submitted). Learning in probabilistic environments.

Hahn, U. & Chater, N. (1998). Similarity and rules: distinct? exhaustive? empirically distinguishable? *Cognition*, 65, 197-230.

Hahn, U., & Nakisa, R.C. (2000). German Inflection: Single or Dual Route? *Cognitive Psychology, 41*, 313-360.

Hauser, M., Chomsky, N., Fitch, W.T. (2002) The faculty of language: What is it, Who has it, and how did it evolve? *Science*, 298 (22), 1569-1579.

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences, 9*, 1-23.

Horning, J.J. (1969). *A study of grammatical inference.* PhD Thesis, Stanford University.

Hurford, J. (2000) The Emergence of Syntax. In C. Knight, M. Studdert-Kennedy, and J. Hurford (Eds.) *The Evolutionary Emergence of Language: Social function and the origins of linguistic form.* Cambridge University Press, pp.219-230.

Hochberg, J., & McAlister, E. (1953). A quantitative approach to figural goodness. *Journal of Experimental Psychology*, 46, 1953, 361--364.

Jusczyk, P.W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences*, 3, 323-328.

Kelly, M.H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, 99, 349-364.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2): 102-110.

Kiss, G. (1973). Grammatical word classes: A learning process and its simulation. *Psychology of Learning and Motivation*, 7: 1-41.

Klavans, J.L., & Resnik, P. (1996). *The Balancing Act: Combining Symbolic and Statistical Approaches to Language.* Cambridge, MA: MIT Press.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems in Information Transmission*, 1, 1-7.

Levin, B. (1993), *English verb classes and alternations.* Chicago: The University of Chicago Press.

Li, M. & Vitányi, P. (1997). *An introduction to Kolmogorov complexity theory and its applications* (2nd edition). Berlin: Springer Verlag

Lord, C. (1979). Don't you fall me down: Children's generalizations regarding cause and transitivity. *Papers and Reports on Child Language Development*, 17. Stanford, CA: Stanford University Department of Linguistics.

Luce, R.D. (1963). Detection and recognition. In R.D. Luce, R.R. bush, & E. Galanter, Eds.) *Handbook of Mathematical Psychology.* New York: Wiley.

MacKay, D.J.C., (1992). Information-based objective functions for active data selection. *Neural Computation*, 4, 589-603.

MacWhinney, B. (1987). The Competition Model. In B. MacWhinney (Ed.), *Mechanisms of language acquisition.* Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B. (1989). Competition and Lexical Categorization. In R. Corrigan, F. Eckman, & M. Noonan (Eds.) *Linguistic categorization,* 195-242. New York: Benjamins.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk.* 3rd Ed. London : Lawrence Erlbaum.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press.

Maratsos, M.P. & Chalkley, M.A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K.E. Nelson (Ed.), *Children's Language Volume 2*, pp.127-214. New York: Gardner Press.

Marchman, V. & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language*, 21(2), 331-366.

Marcus, G.F. (1999). Do infants learn grammar with algebra or statistics? *Science*, 284, 436-437.

Marcus, G.F. (2001). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press.

Marcus, G.F., & Berent, I. (2003). Are there limits to statistical learning? *Science*, 300, 52-53.

Marcus, G.F., Vijayan, S., Bandi Rao, S., Vishton, P.M. (1999). Rule Learning by Seven-Month-Old Infants. *Science,* 283: 77-80.

McClelland, J.L. (1998). Connectionist models and Bayesian inference. In M.Oaksford, & N. Chater (Eds.) *Rational models of cognition*. Oxford: Oxford University Press.

McClelland, J.L., & Plaut, D.C. (1999) Does generalisation in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences*, 3, 166-168.

Miller, G.A. (1967). Project Grammarama, in *The Psychology of Communication: Seven Essays*. Baltimore: Penguin.

Mintz, T.H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition* , 30, 678-686.

Mintz, T. H. (in press). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition.*

Mintz, T.E., Newport, E., & Bever, T.G. (1995). Distributional regularities in speech to young children. In *Proceedings of NELS*, 25, 43-54.

Mintz, T.H. (2002). Category Induction from Distributional Cues in an Artificial Language. *Memory and Cognition*, 30, 678-686.

Mintz, T.H., Newport, E.L., & Bever, T.G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26, 393-425.

Monaghan, P., Chater, N. & Christiansen, M.H. (submitted). The differential contribution of phonological and distributional cues in grammatical categorisation.

Monaghan, P., Chater, N., & Onnis, L. (in preparation). Optimal data selection in sequential AGL learning.

Morgan, J.L., Newport, E. (1981). The role of constituent structure in the induction of an artificial language. *Journal of verbal learning and verbal behavior*, 20: 67-85.

Newport, E., & Aslin, R.N. (2000). Innately constrained learning: Blending old and new approaches to language acquisition. In S.C. Howell, A. Fish, & T. Keith-Lucas (Eds.) *Proceedings of the 24th Annual Boston University Conference on Language Development.*

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review,* 101(4), 608-631.

Onnis, L., Roberts, M., & Chater, N. (2002) Simplicity: A cure for overgeneralizations in language acquisition? In W.D. Gray & C.D. Shunn, (Eds.) *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, London: LEA.

Onnis, L., Christiansen, M.H., Chater, N., & Gómez, R. (2003). Reduction of Uncertainty in Human Sequential Learning: Evidence from Artificial Grammar Learning.. *Proceedings of the 25th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Onnis, L., Destrebecq, A., Christiansen, M., Chater, N., & Cleeremans, A. (submitted). Learning nonadjacent dependencies: A graded, distributed account.

Onnis, L., Gómez, R., Christiansen, M.H., & Chater, N. (in preparation). Detecting invariant structure and generalising under conditions of variability.

Onnis, L., Monaghan, P., Chater, N. & Richmond, K. (submitted). Phonology impacts segmentation and generalisation in speech processing.

Peña, M., Bonatti, L., Nespor, M., Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, 298, 604-607.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General*, 119, 264-275.

Pine, J.M., & Lieven, E.V.M. (1997). Slot-and-frame patterns and the development of the determiner category. *Applied Psycholinguistics*, 18, 123-138.

Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, MA: MIT Press.

Pinker, S. (1995). *The language instinct*. Harmondsworth: Penguin.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Basic Books.

Pizzuto E., & Caselli, M.C. (1992). The acquisition of Italian morphology: implications for models of language development. *Journal of Child Language*, 19, 491-557.

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.

Plunkett, K. & Juola P. (1999). A connectionist model of english past tense and plural morphology. *Cognitive Science*, 23, (4), 463-490.

Pothos, E., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science*, 26, 303-343.

Prasada, S., & Pinker, S. (1993). Generalizations of regular and irregular morphology. *Language and Cognitive Processes*, 8, 1-56.

Quinlan, J.R. & Rivest, R. (1989). Inferring decision trees using the minimum description length principle. *Information and Computation*, 80, 227-248.

Reber, A.S. (1969). Transfer of syntactic structure in synthetic languages. *Journal of Experimental Psychology*, 81, 115-119.

Redington, M., & Chater, N. (2002). Knowledge representation and transfer in artificial grammar learning (AGL). In R. M. French & A. Cleeremans (Eds.) *Implicit learning and consciousness: an empirical, philosophical, and computational consensus in the making*. Psychology Press.

Redington, M., Chater, N., and Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4): 425-469.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society*, Series B, 49, 223-239.

Rissanen, J. (1989). *Stochastic complexity and statistical inquiry*. Singapore: World Scientific.

Roberts, M., Onnis, L. & Chater (in press). Acquisition and evolution of languages: Two puzzles for the price of one. To appear in: Tallerman, M., *Pre-requisites for language evolution.*

Rumelhart, D.E., & McClelland, J.L. (1986). On learning the past tense of English verbs. In J.L. McClelland, D.E. Rumelhart and the PDP Research Group *Parallel distributed processing: Explorations in the microstructure of cognition. Vol 2: Psychological and biological models*, pp.216-271. Cambridge, MA: MIT Press.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35, 606-621.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science,* 274, 1926-1928.

Saffran, J.R., Johnson, E.K., Aslin, R.N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition,* 70, 27-52.

Schvaneveldt, R.W., & Gómez, R.L. (1998). Attention and probabilistic sequence learning. *Psychological Research,* 61, 175-190.

Seidenberg, M. S., & McClelland, J.L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96, 523-568.

Seidenberg, M.S., MacDonald, M.C., & Saffran, J.R. (2002). Does grammar start where statistics stop? *Science*, 298, 553-554.

Seidenberg, M.S., MacDonald, M.C., & Saffran, J.R. (2003). Response to Marcus and Berent. *Science*, 300, 53.

Servan-Schreiber, D., Cleeremans, A. & McClelland, J.L. (1991). Graded State Machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning*, *7*, 161-193.

Servan-Schreiber, E., & Anderson, J.R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 592-608.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30, 50-64.

Shannon, C.E. (1948). A Mathematical Theory of Communication, *Bell System Technical Journal, 27,* 379-423 and 623-656.

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.

Stolcke, A. (1994). *Bayesian Learning of Probabilistic Language Models.* Doctoral dissertation. Department of Electrical Engineering and Computer Science. University of California at Berkeley.

Teal, T.K. and Taylor, C.E. (2000). Effects of Compression on Language Evolution. *Artificial Life*, 6 (2): 129-143.

Tomasello, M. (2000). First steps towards a usage-based theory of language acquisition. *Cognitive Linguistics*, 11, 61-82.

Tomasello, M. (2000). The item based nature of children's early syntactic development. *Trends in Cognitive Sciences,* 4, 156-163.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition.* Harvard University Press.

Van der Helm, P.A., & Leeuwenberg, E.L.J. (1996). Goodness of visual regularities: A non-transformational approach. *Psychological Review*, 103 (3), 429-456.

Vokey, J.R., & Brooks, L.R. (1992). Salience of item knowledge in learning artificial grammar. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 328-344.

Wallace, C.S., & Freeman, P.R. (1987). Estimation and inference by compact coding. *Journal of the Royal Statistical Society Series* B, 49 (3), 240-265.

Wolff, J. (1982). Language acquisition, data compression and generalization. *Language & Communication*, 2, 57-89, 1982.

Wolff, J. (1991). *Towards a Theory of Cognition and Computing*. Chichester: Ellis Horwood.

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley, Reading, MA.

Zuidema, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In S. Becker, S. Thrun, and K. Obermayer (Eds.) *Advances in Neural Information Processing Systems* 15, Cambridge, MA: MIT Press.

APPENDIX A

For the language to be probabilistically generated and understood, it was necessary to assign several sets of probabilities. I took the probability of any given linguistic construction to be a function of the frequency of its components. Thus constructions comprised of highly frequent words are taken to be much more probable than those comprised of low frequency words. This was done by applying Zipf's law (Zipf, 1948) which states that the frequency of any word is given as the inverse of its rank. This distribution is frequently encountered in natural languages (see, e.g. Bell *et al.*, 1990)

Initially all words, As and Bs, were ranked arbitrarily. Subsequently all possible sentences allowable under the production rules were generated, minus any specified in the exceptions element. The result is illustrated in Figure 2.1. Each word could occur in a number of distributional contexts, with different probabilities for occurrence in each.

w_1		w_2	***Rank 1***	***Rank 2***	***Rank 3***	***Rank 4***
Rank 1	a_3		b_2	b_4	b_1	b_3
Rank 2	a_1		b_2	b_4	b_1	b_3
Rank 3	b_2		a_3	a_1	a_2	a_4
Rank 4	a_2		b_2	b_4	b_1	b_3
Rank 5	b_4		a_3	a_1	a_2	a_4
Rank 6	b_1		a_3	a_1	a_2	a_4
Rank 7	a_4		b_2	b_4	b_1	b_3
Rank 8	b_3		a_3	a_1	a_2	a_4

Figure 2.1. A completely regular hypothesis grammar. The left hand columns show a frequency ranking for all As and Bs as the first word of any sentence. The right hand columns show the frequency rankings of As and Bs as the the second word of any sentence given the first word. For example, word b_2 was the most likely to occur in position two with a_3 in position one, but the third most likely to occur in position one. No exceptions are specified in [1] so all sentences were allowable.

The probability of a word occurring in a particular distributional context is given as:

$$p = \frac{f}{\sum f} \qquad [4]$$

where p is the probability of a word, f is the frequency of that word and $\sum f$ are the frequencies of the n words in the distribution. Any sentence, involves two probabilities $p_{(w1)}$ and $p_{(w2|w1)}$ where $p_{(w1)}$ is the probability of the first word and $p_{(w2|w1)}$ is the probability of the second word in the distributional context of the first word (see Figure. 2.1). $p_{(w1)}$ is given by equation [2] with $\sum f$ operating over all eight words. For $p_{(w2|w1)}$, $\sum f$ operates over the distribution of possible second words associated with w_1. With no exceptions specified there were always four possible second words (Figure 2.1). If exceptions were specified, however, the number of possible second words would vary between first words (Fig. 2.2).

Once a table such as those in Figures 2.1 and 2.2 had been set up, samples of language were produced by generating random probabilities to select the first word of the sentence and then the second word given the first.

w_1			w_2	*Rank 1*	*Rank 2*	*Rank 3*	*Rank 4*
Rank 1	a_3			b_2	b_4	b_1	b_3
Rank 2	a_1			b_2	b_4	b_1	b_3
Rank 3	b_2			a_3	a_1	a_4	
Rank 4	a_2			b_2	b_4	b_1	b_3
Rank 5	b_4			a_3	a_1	a_4	
Rank 6	b_1			a_3	a_1	a_4	
Rank 7	a_4			b_2	b_4	b_1	b_3
Rank 8	b_3			a_3	a_1	a_4	
Exceptions:	b_1,a_2	b_2,a_2	b_3,a_2	b_4,a_2			

Figure 2.2. A hypothesis grammar containing exceptions. In this specification, a_2 can only appear in the first word position. A number of sentences are therefore specified as exceptions. This alters the number of possible second words following some first words.

It was possible to specify both data and hypotheses in exactly the same way. All learners entertained one initial hypothesis. This was the completely regular hypothesis expressed in [3]. The only difference between this and later hypotheses was the number of exceptions specified. The code length necessary to specify the syntactic categories A and B and the production rules were identical for every hypothesis and therefore need not be considered. The only hypothetical element that needed to be specified was the final, exceptions, element. This element, when it was not empty, consisted of a set of sentences of exactly the same form as those generated as samples of the language. The code length necessary to specify an exception was therefore exactly the same as the code length necessary to specify that sentence were it to be encountered as data. Following [1], the code length necessary to specify a sentence w_1,w_2 is given as:

$$bits_{(w1,w2)} = Log_2\left(\frac{1}{p_{(w1)} \cdot p_{(w2|w1)}}\right) \qquad [5]$$

where $bits_{(w_1,w_2)}$ is the number of bits necessary to specify sentence w_1,w_2, is the probability of w_1 and is the probability of w_2 given w_1. These values are found using [4]. In the event that the second word was unknown given the first, i.e. that the sentence was disallowed under the hypothesis, the code length necessary to specify it was:

$$bits_{(w1,w2)} = Log_2\left(\frac{1}{p_{(w1)} \cdot p_{(w2)}}\right) \qquad [6]$$

where $P_{(w_2)}$ is the probability of w_2 irrespective of w_1, as if it were a first word. The second word was thus coded as if it were one of eight ranked possibilities making the overall probability of the sentence lower than if it were allowable and increasing the code length. In this way hypotheses that posited spurious exceptions were punished with longer data code lengths when those exceptions were encountered.

As mentioned above, each learner agent began by entertaining a single completely regular hypothesis without any exceptions. Initially, therefore, all data was coded under one hypothesis only. As more hypotheses emerged they ran in parallel with previous ones so that data coded under all hypotheses simultaneously. Each new hypothesis was a clone of its immediate predecessor with the addition of one exception. Thus the initial hypothesis contained no exceptions, the second contained one, the third two and so on. A new exception was postulated when a particular construction had never been heard and an MDL-derived parameter, [7] was satisfied. A derivation is given at the end of this appendix:

$$N / \left(\frac{\log_2(1/p)}{p}\right) \qquad [7]$$

where N is the total number of sentences heard so far and p is the probability of a particular sentence. This parameter merits some discussion.

A learner's decision to posit a particular sentence as an exception is dependent on two data: the total number of sentences heard and the number of times that the sentence in question has been heard. How these are combined to determine the precise point at which an exception is posited is to some extent arbitrary. For simplicity, I will only consider the case in which no sentence is ever posited as an exception if it has been encountered in the data. The critical value that determines when a particular sentence is posited as an exception is therefore the number of sentence that have been heard. Two normative criteria for this threshold exist: on the one hand it should not be so low that the learner concludes there is an exception when in fact none exists; on the other, the learner should not fail to spot genuine exceptions after a exposure to a reasonable amount of data. The consequences of failure to meet either of these criteria can be seen in both cognitive and linguistic terms. Both will result in longer codelengths: the former will incur long data codes when it encounters the sentences that it has specified as exceptions; the latter will incur long data codes that it could reasonably have avoided by specifying exceptions earlier. Linguistically, in the former case the learner will have legitimate sentences pruned from its productive repertoire; in the latter it will continue to produce illegitimate sentences for longer than necessary.

In these simulations not all sentences were equally probable. Less probable (and absent) sentences should require more language to be encountered before they could be considered exceptions. This was taken into account by making use of a general derivation (not specific to these simulations) based on the premise

that an exception should be postulated at the point at which the investment of bits necessary to specify it would have been recouped had it been postulated before any language was heard.

Suppose that a learner wants to know whether to consider sentence x as an exception, where is $p_{(x)}$ the probability of x. If it is postulated as an exception, we can increase the probability of the other sentences that have not been ruled out. These probabilities used to sum to $1-p_{(x)}$ but with x as an exception they sum to 1. The most neutral way to rescale these probabilities is to multiply them all by the same factor $\frac{1}{1-p_{(x)}}$. This increase means that the code for each item reduces by $\log_2\left(\frac{1}{1-p_{(x)}}\right)$ (See [1] in the main text). Thus if the learner hears a corpus of N sentences, never encountering x and having postulated x as an exception, it will make a saving of $N\log_2\left(\frac{1}{1-p_{(x)}}\right)$ over the whole corpus. Thus x may be postulated as an exception when this saving exceeds the cost of specifying x as an exception:

$$\log_2\left(\frac{1}{p_{(x)}}\right) > N\log_2\left(\frac{1}{1-p_{(x)}}\right)$$

If we assume that the probability of any particular sentence is small (i.e. near 0), a Taylor expansion gives that $\log_2\left(\frac{1}{1-p_{(x)}}\right)$ approximately equals $p_{(x)}$. From this we can conclude [7]