This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Cognitive biases in competition: Innovation and the evolution of language structure

Annie Holtz
Lay summary

There is a huge amount of diversity across the over 7,000 languages currently used in the world. This diversity exists at every level of the language system, from the vocabulary and speech sounds/signs used to express individual words, to the abstract organisation of rules that govern grammatical processes in different languages. For example, a phrase like *seven red skirts* and the phrase with the same meaning in Swedish, *sju röda kjolar*, use different words and different speech sounds within those words. Despite this variety, there are repeated patterns across the languages of the world, showing that some language structures are more common than other possible alternatives. In fact, the sentences above exhibit a striking similarity in that the order of the words in the sentences are the same. In fact, the order used in these two phrases to express number, colour and item is one of the most common orders found across all languages. These recurring patterns do not seem to simply be a by-product of historical relationships between languages, nor a random selection from all the possible patterns that could exist in languages. So what is the root cause of these similarities?

One possible cause of these similarities is that human languages are restricted by features of the human mind. These features are often referred to by researchers as cognitive biases, and they bias individuals to favour or disfavour certain language features. This thesis focuses on examining the impact of a set of these cognitive biases on typological language structure related to word order. Using a combination of data comparison across spoken and signed languages, as well as a review of previous experimental literature on the impact of cognitive biases on language structure, I identify a new division between different types of cognitive biases. This division is based on the fact that some of these preferences seem to influence the way we construct and interpret whole linguistic systems, whereas others influence the way we express individual linguistic meanings without reference to the wider language system. Based on this division, I develop a new theory which allows us to make predictions about when we might expect to see each type of bias influencing people’s interaction with linguistic data. These predictions are
tested in a series of experiments where participants are either trained on a miniature gestural language system, or asked to judge the suitability of various novel linguistic expressions in isolation.

The results of these experiments show that two different bias types are mainly active in separate language contexts. I construct a scale where these language contexts are placed at each end, and show how other types of contexts can be ordered between them based on how much language learning versus language innovation is required to complete tasks in each context. My final conclusion is that there are tasks in the centre of this scale that combine the need to learn a language and to innovate new linguistic expression, and that in these tasks we can observe some influence of both types of biases at the same time. This combined effect of biases that influence language systems and biases that influence items within those systems helps explain several patterns of similarity that we see across languages, and provide interesting insights into how cognitive biases may have influenced the progression of language evolution.
Abstract

The cognitive system that underpins our capacity for language has the potential to systematically constrain the type of language structures we see represented across the world’s languages. This thesis follows a research tradition which conceptualises these constraints as probabilistic cognitive biases that influence different aspects of our interaction with linguistic data, such as language learning, communication, and even language creation. I explore how such biases give rise to cross-linguistic regularities by examining previous experimental evidence across phonology and syntax, and combining those insights with typological data from both spoken and signed languages to show how the impact of cognitive biases can be observed within individuals and on a population-wide scale. This review of previous research motivates a novel division of cognitive biases into two groups, system-wide biases that influence the way we organise and structure language systems, and category-specific biases that influence our interaction with specific linguistic items within those systems. These biases are then mapped onto a theoretical construct I call the Scale of Innovation, which orders linguistic contexts and tasks along a continuum based on how much an individual in that context needs to innovate to complete a given linguistic task. Combined, this bias division and the Scale of Innovation allows me to make predictions regarding the contexts in which we can expect different biases to be active, as well as their ability to affect cross-linguistic structure and the progression of language evolution. These predictions are tested in a series of artificial language learning tasks and silent gesture experiments which are purposefully manipulated to represent different regions along the Scale of Innovation.

The results of the thesis show how the influence of category-specific and system-wide biases pattern across the Scale of Innovation. Evidence for the effect of system-wide biases are mainly found in language learning contexts, which are low on the scale as they require little innovation, whereas category-specific biases are mainly found in language improvisation tasks, which are high on the scale as they require high amounts of innovation. Furthermore, the biases are shown
to have simultaneous and (sometimes) competing effects in the central regions of the scale, in contexts that combine both learning and improvisation. As such, the thesis presents three main contributions; (i) it outlines the novel division between category-specific and system-wide cognitive biases, (ii) it establishes the Scale of Innovation as a useful conceptual tool in studies of cross-linguistic regularities and language evolution, and (iii) it provides the missing link showing how category-specific biases can influence wider typological structure by showing that their effects on behaviour are not limited to pure improvisation tasks.
Acknowledgements

This is it, the beginning of the end. We finally made it. All mentioned below have played an integral part in my PhD journey. Know that I have a thousand more words of thanks for each one of you, but this thing is already very long so what you read here will have to suffice.

Thank you to my supervisors, Jennifer Culbertson and Simon Kirby, for taking a chance on a bright-eyed undergraduate student when I decided to apply for my PhD. You were both sources of inspiration to me long before you knew who I was, and you continued to inspire me throughout my PhD. Thank you for pushing me to try things I never would have dared to do on my own, and for recognising that my talents went beyond what I produced in my research. Your continued ability to juggle so many projects and responsibilities in the middle of a pandemic boggles my mind. The CLE is lucky to have two such dedicated and inquisitive minds at its helm. This thanks also extends to Alexander Martin, who helped supervise me for a brief period at the start of my PhD.

There is no way I could have produced the thesis you see before you without the many friends and colleagues I had within the CLE. Marc Meisezahl, my “PhD twin”, has shared in my every academic trial and triumph and I have loved watching us both grow as researchers. Fang Wang was the calm we all needed during the storms of supervision. Aislinn Keogh and Elizabeth Pankratz should really both get credit as co-supervisors for our many brainstorming sessions and how generously they shared their expertise with me. Every conversation with Henry Conklin broadened my horizons and taught me something new (I am so glad you came to town too).

I also wish to thank the other CLE members, past and present, who have been emotional support, givers of advice, organisers of fun and founts of wisdom. Special mentions go out to Maisy Hallam, Shira Tal, Yevgen Matusevych, and Yasamin Motamed but there are literally too many of you to list so just import the whole CLE “members” and “past members” web-page here and that should about cover it!
The University of Edinburgh also gave me several other important friendships during both my undergraduate and my postgraduate degree. I would be lost and hungry without the academic insight and cake provided by my wonderful friend Nina Markl. She, Bran Papineau and Emma Kouhi make up the most supportive and quotable group chat of all time. Edinburgh also introduced me to the brilliant Lily Chase and her wonderful partner Léa Davoine, Zoom calls with them during the pandemic were like a balm to the soul. Thank you to Matt King, for many thought-provoking conversations at various cafes across town. Then there is Gilly Marchini and Carine Abraham who both just bring sunshine energy into anyone’s life.

But wait! I do not just have academic friends, I promise. For eleven years Sophie Woolford and Ottília Westerlund have been my biggest cheerleaders. The number of linguistics rants they have been subjected to in that time is fairly embarrassing, but hey at least it paid off in the end. Thank you for long chats about life, fun nights of distraction, and wonderful weekends in the mountains. The same thanks extend to other friends and loved ones who, while far away, have managed to support me in so many ways. So thank you to Milena Asklo, Annabel Reid, Rosanna Langan, Nina Wahlberg, and Mats Wahlberg for being there when I need you.

My relatives have also been an incredible source of comfort throughout this process. Thank you to Gunilla, Ulf, Gun, Sebastian (+Bailey), Ingrid, Agneta, Krille, Tom, Annika, John, Petronella, Stella, Axel, Inger, Harold and many others in my sprawling family tree for your steadfastness and interest in my odd choice of career.

Finally, I want to thank my family, Caroline, Jan and Tina, and my partner, Niklas. What can I even say? I always have the words (usually too many) for everything, but I truly do not know how to thank you for all your faith, all your support, and all your encouragement. This thesis is a labour of love, and I know exactly where all that love came from. Tack till er alla.
Contents

Lay summary 4

Abstract 6

Acknowledgements 8

1 The influence of cognitive biases on language structure and be-
behaviour 21
  1.1 Introduction ........................................... 21
  1.2 Chapter outline ....................................... 24
  1.3 Typological regularities .............................. 25
      1.3.1 Language universals ........................... 27
      1.3.2 Explanations of typological regularities ...... 29
  1.4 Cognitive biases and linguistic behaviour .......... 32
      1.4.1 Simplicity phenomena .......................... 35
      1.4.2 Naturalness phenomena ......................... 41
      1.4.3 Category-specific and system-wide biases ...... 48
  1.5 Scale of Innovation ................................... 50
      1.5.1 Improvising ...................................... 51
      1.5.2 Extrapolating .................................... 55
      1.5.3 Generalising .................................... 58
      1.5.4 Regularising .................................... 60
      1.5.5 Memorising ...................................... 63
      1.5.6 Connecting the scale and cognitive biases ...... 65
  1.6 Case study: competition between system-wide and category-specific biases in nominal word order .... 66
      1.6.1 Order of noun and adjective .................... 68
      1.6.2 Order of noun and genitive ..................... 69
      1.6.3 Biases in competition ........................... 73
  1.7 Thesis outline ......................................... 74
2 With or without a system: How category-specific and system-wide cognitive biases shape word order

2.1 Preface: Chapter 2

2.2 Holtz et al. (under revision): author contributions

2.3 Abstract

2.4 Introduction

2.4.1 Category-specific and system-wide biases

2.4.2 Different biases in different contexts

2.5 Experiment 1

2.5.1 Methods

2.5.2 Coding

2.5.3 Results

2.5.4 Experiment 1: Discussion

2.6 Experiment 2

2.6.1 Methods

2.6.2 Results

2.6.3 Experiment 2: Discussion

2.6.4 General Discussion

2.7 Conclusion

References

Chapter appendices

2.A Outline

2.B Stimuli design

2.B.1 Image stimuli creation

2.B.2 Video stimuli creation

2.B.3 Stimuli motivation

2.C Consent and instructions

2.C.1 Experiment 1: Consent info

2.C.2 Experiment 1: Instructions

2.C.3 Experiment 2: Consent info

2.C.4 Experiment 2: Instructions

2.D Language data

2.E Translation data

2.F Motivation data

2.G Monte Carlo simulation
3 Item-specific and system-wide preferences in competition: Effects on noun phrase harmony 145

3.1 Preface: Chapter 3 ........................................... 145
3.2 Holtz et al. (under review): author contributions .............. 147
3.3 Abstract ...................................................... 148
3.4 Introduction ................................................... 148
3.5 Experiment .................................................... 152
  3.5.1 Methods .................................................. 153
  3.5.2 Results ................................................... 156
3.6 Discussion .................................................... 162
3.7 Conclusion .................................................... 166
References ......................................................... 168

Chapter appendices 173

3.A Outline ....................................................... 173
3.B Consent info .................................................. 173
3.C Instructions .................................................... 175
3.D Entropy CIs comparison ..................................... 176
3.E Language data ................................................ 176
3.F Translation data ............................................... 178
3.G Monte Carlo simulation ...................................... 178
3.H Markov chain analysis ....................................... 179

4 Does degree of innovation modulate the influence of cognitive biases on language structure? 187

4.1 Preface: Chapter 4 ........................................... 187
4.2 Holtz et al. (under review): author contributions .............. 189
4.3 Abstract ...................................................... 190
4.4 Introduction ................................................... 190
  4.4.1 Scale of Innovation ..................................... 194
  4.4.2 Category-specific biases across the scale ................. 196
4.5 Experiment 1 .................................................. 198
  4.5.1 Method .................................................... 198
  4.5.2 Results ................................................... 202
  4.5.3 Discussion ................................................ 210
4.6 Experiment 2 .................................................. 211
  4.6.1 Method .................................................... 212
  4.6.2 Results ................................................... 212
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Examples of phonological alternations involving different numbers of features</td>
<td>40</td>
</tr>
<tr>
<td>1.2</td>
<td>Distribution of possible NP orders and the hierarchical structure of the NP</td>
<td>47</td>
</tr>
<tr>
<td>1.3</td>
<td>Visual representation of the Scale of Innovation</td>
<td>51</td>
</tr>
<tr>
<td>2.1</td>
<td>Distribution of the 24 possible NP orders showing homomorphic prevalence</td>
<td>86</td>
</tr>
<tr>
<td>2.2</td>
<td>Sample image stimuli from the adjective condition and the genitive condition</td>
<td>91</td>
</tr>
<tr>
<td>2.3</td>
<td>Illustrative examples of the main trials in Experiment 1</td>
<td>93</td>
</tr>
<tr>
<td>2.4</td>
<td>Comparison of proportions of postnominal orders across typology and in the experimental results</td>
<td>95</td>
</tr>
<tr>
<td>2.5</td>
<td>Summary of strength of preference data based on slider trial in Experiment 1</td>
<td>97</td>
</tr>
<tr>
<td>2.6</td>
<td>Summary of majority order selection across conditions for Experiment 2</td>
<td>103</td>
</tr>
<tr>
<td>2.7</td>
<td>Summary of change in overall entropy across conditions for Experiment 2</td>
<td>105</td>
</tr>
<tr>
<td>2.8</td>
<td>Summary of natural order selection across conditions for Experiment 2</td>
<td>107</td>
</tr>
<tr>
<td>2.9</td>
<td>Summary of change in mutual information based on meaning across conditions for Experiment 2</td>
<td>108</td>
</tr>
<tr>
<td>2.B.1</td>
<td>All image stimuli for genitive meanings</td>
<td>124</td>
</tr>
<tr>
<td>2.B.2</td>
<td>All image stimuli for adjective meanings</td>
<td>125</td>
</tr>
<tr>
<td>2.B.3</td>
<td>Sample stills from gesture videos showing sequences of two gestures for both adjective and genitive meanings</td>
<td>126</td>
</tr>
<tr>
<td>2.G.1</td>
<td>Distribution of simulated means of change in overall entropy across conditions for Experiment 2 in Chapter 2</td>
<td>141</td>
</tr>
</tbody>
</table>
2.G.2 Distribution of simulated means of change in overall entropy for each condition for Experiment 2 in Chapter 2. 142
2.G.3 Distribution of simulated means of change in MI based on head noun across conditions for Experiment 2 in Chapter 2. 142
2.G.4 Distribution on simulated means of change in MI based on head noun for each condition for Experiment 2 in Chapter 2. 143

3.5.1 Sample stimuli for adjective meanings and genitive meanings. 154
3.5.2 Stills of gestures used to denote the meaning striped book 154
3.5.3 Proportion of postnominal orders selected by each participant for both dependent types across all conditions 156
3.5.4 Proportion of test trials where participants selected the majority input order for each condition and dependent type 157
3.5.5 Proportion of test trials where participants selected the natural order (i.e., prenominal for genitives, postnominal for adjectives) for each condition and dependent type 158
3.5.6 Summary of mean change in overall entropy between input and output across conditions 159
3.5.7 Summary of mean change in mutual information input and output across conditions based on dependent type 162
3.5.8 Summary of mean change in mutual information input and output across conditions based on head noun 163
3.G.1 Distribution of simulated means of change in overall entropy across conditions for the experiment in Chapter 3 179
3.G.2 Distribution based on simulated means of change in overall entropy for each condition for the experiment in Chapter 3 180
3.G.3 Distribution of simulated means of change in MI based on head noun across conditions for the experiment in Chapter 3 181
3.G.4 Distribution of simulated means of change in MI based on head noun for each condition for the experiment in Chapter 3 182
3.G.5 Distribution of simulated means of change in MI based on dependent type across conditions for the experiment in Chapter 3 183
3.G.6 Distribution of simulated means of change in MI based on dependent type for each condition for the experiment in Chapter 3 184
3.H.1 Representation of initialising transition matrix underlying Markov chain process for Chapter 3 185
3.H.2 Stationary distribution across the four language types for Chapter 3 .......................................................... 185

4.4.1 Scale of Innovation .................................................. 195
4.5.1 Sample familiarisation trials showing both adjective and genitive meanings. ........................................ 201
4.5.2 Sample trial from the testing phase for the target meaning spotted hat. ...................................................... 202
4.5.3 Proportion of postnominal orders selected by each participant for both dependent types across all conditions in Experiment 1. ... 204
4.5.4 Proportion of test trials where participants selected the majority input order for each condition and dependent type in Experiment 1 205
4.5.5 Proportion of test trials where participants selected the natural order for each condition and dependent type in Experiment 1 ... 206
4.5.6 Mean change in entropy between input and output, showing participants tend to harmonise across all conditions in Experiment 1 .......................................................... 207
4.5.7 Mean change in MI based on dependent type, showing reliable change in MI across all conditions except the unnatural condition in Experiment 1 ...................................................... 209
4.5.8 Mean change in MI based on head nouns, showing reliable change in MI across all conditions in Experiment 1 .......................................................... 210
4.6.1 Proportion of postnominal orders selected by each participant for both dependent types across all conditions in Experiment 2 ... 214
4.6.2 Proportion of test trials where participants selected the majority input order for each condition and dependent type in Experiment 2 2215
4.6.3 Proportion of test trials where participants selected the natural order for each condition and dependent type in Experiment 2 ... 216
4.6.4 Mean change in entropy between input and output, showing participants tend to harmonise in the natural and unnatural condition, but not in the majority prenominal or majority postnominal condition in Experiment 2 .......................................................... 217
4.6.5 Mean change in MI based on dependent type in Experiment 2, showing reliable change in MI for the two majority order conditions, but not the natural or unnatural condition .................... 218
4.6.6 Mean change in MI based on head nouns in Experiment 2, showing slight increase in MI across all conditions ...................... 219
4.F.1 Distribution of simulated means of change in overall entropy
across conditions for Experiment 1 in Chapter 4 . . . . . . . . .
4.F.2 Distribution based on simulated means of change in overall entropy for each condition for Experiment 1 in Chapter 4 . . . . .
4.F.3 Distribution of simulated means of change in MI based on head
noun across conditions for Experiment 1 in Chapter 4. . . . . . .
4.F.4 Distribution of simulated means of change in MI based on head
noun for each condition for Experiment 1 in Chapter 4 . . . . .
4.F.5 Distribution of simulated means of change in MI based on dependent type across conditions for Experiment 1 in Chapter 4. . . .
4.F.6 Distribution of simulated means of change in MI based on dependent type for each condition for Experiment 1 in Chapter 4 . . .
4.F.7 Distribution of simulated means of change in overall entropy
across conditions for Experiment 2 in Chapter 4 . . . . . . . . .
4.F.8 Distribution based on simulated means of change in overall entropy for each condition for Experiment 2 in Chapter 4 . . . . .
4.F.9 Distribution of simulated means of change in MI based on head
noun across conditions for Experiment 2 in Chapter 4. . . . . . .
4.F.10 Distribution of simulated means of change in MI based on head
noun for each condition for Experiment 2 in Chapter 4 . . . . .
4.F.11 Distribution of simulated means of change in MI based on dependent type across conditions for Experiment 2 in Chapter 4. . . .
4.F.12 Distribution of simulated means of change in MI based on dependent type for each condition for Experiment 2 in Chapter 4 . . .

243
244
244
245
245
246
247
248
248
249
249
250

5.2.1 Thesis experiments mapped to the associated regions along the
Scale of Innovation . . . . . . . . . . . . . . . . . . . . . . . . . 257
5.3.1 Scale of Innovation mapped to proposed progression of language
evolution . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 264

16


# List of Tables

1.1 Illustrative example of rules in a harmonic and a non-harmonic language ........................................ 36
1.2 Typological counts of the order of nouns with adjectives in spoken and signed languages .................. 69
1.3 Typological counts of the order of nouns with genitives in spoken and signed languages ................... 69
1.4 Typological counts of the order of nouns with numerals in spoken and signed languages .................. 74
1.5 Typological counts of the order of nouns with adjectives and numerals in spoken and signed languages 75
1.6 Typological counts of the order of nouns with adjectives and genitives in spoken and signed languages 75
2.1 Order of adjectives in relation to nouns showing postnominal preference ..................................... 88
2.2 Order of genitives in relation to nouns showing prenominal preference ........................................... 88
2.3 Order of adjectives and genitives in relation to nouns showing that the two most common orders is the postnominal harmonic order and the Gen-N-Adj order 88
2.4 Percentage of prenominal and postnominal gesture orders in input per condition .......................... 100
2.5 Comparison of experimental means and simulated means and resulting z-scores based on change in entropy across conditions for Experiment 2 ........................................ 105
2.6 Comparison of mean entropy change between conditions for Experiment 2 .................................. 106
2.7 Comparison of experimental means and simulated means and resulting z-scores based on change in mutual information across conditions for Experiment 2 ........................... 108
3.H.1 Number of participants per condition in the experiment in Chapter 3 and number of participants that could be classified for the Markov chain analysis .................................................. 186

4.4.1 Typological counts of the order of nouns with adjective and noun with genitive, in spoken and signed languages ............... 198

4.4.2 Typological counts of the order of noun, genitive and adjective, based on spoken language typological data and signed language typological data ...................................................... 198

4.5.1 Percentage of prenominal and postnominal gesture orders for each dependent type in the training data across the four experimental conditions. .................................................. 199

4.5.2 Experimental means, simulated means and resulting z-scores for change in overall entropy in Experiment 1 ................. 208

4.5.3 Experimental means, simulated means and resulting z-scores for change in MI based on dependent type in Experiment 1 .... 209

4.5.4 Experimental means, simulated means and resulting Z-scores for change in MI based on head noun in Experiment 1 ........ 211

4.6.1 Experimental means, simulated means and resulting z-scores for change in overall entropy in Experiment 2 ................. 216

4.6.2 Experimental means, simulated means and resulting Z-scores for change in MI based on dependent type in Experiment 2 .... 218

4.6.3 Experimental means, simulated means and resulting Z-scores for change in MI based on head noun in Experiment 2 ........ 220

4.7.1 Mapping of coding for variables nat_unnat and harm_non across both Experiments 1 and 2 for experiment comparison .... 221

4.D.1 Comparison of overall mean entropy change between conditions for Experiment 1 in Chapter 4 .................................... 239

4.D.2 Comparison of overall mean entropy change between conditions for Experiment 2 in Chapter 4 .................................... 240

4.E.1 Languages (other than English) spoken by participants whose data was included in Experiment 1 in Chapter 4 .............. 240

4.E.2 Languages (other than English) spoken by participants whose data was included in Experiment 2 in Chapter 4 .............. 241
Chapter 1

The influence of cognitive biases on language structure and behaviour

1.1 Introduction

Humans live in complex social groups where we encounter a variety of practical and social problems every day. Over the course of human history we have invented many technological solutions to these everyday problems. For example, we have developed vehicles that allow us to travel long distances in a short period of time with minimal physical effort, we have created hand-held devices that give us access to a wealth of information with just a few quick internet searches, and discovered medicines and treatments that make previously life-threatening conditions easily curable. Our ability to achieve all of these advances rely on one fundamental human skill, our ability to communicate with each other using language. Language allows us to efficiently pass on knowledge from one individual to another with high fidelity, either through interactional social transmission, or using written records (Boyd et al., 2011). This ability is the foundation for the development of a wider process, known as cumulative cultural evolution (Mesoudi & Thornton, 2018). Cumulative cultural evolution means that new generations can build upon the knowledge base given to them by the previous generation, resulting in faster development of more advanced and optimal technologies (Boyd et al., 2013; Derex & Mesoudi, 2020; Henrich, 2016; Kirby et al., 2008).

While language is crucial for the process of cultural evolution, it is also itself a product, or artefact, created by cumulative cultural evolution as a solution to the problem of communication (Sterelny, 2016).\footnote{There is a debate between researchers who believe that language evolved as a response to the need to communicate (Carruthers, 2002; Jackendoff, 2002; Levinson, 2019; Scott-Phillips, 2014) and those who believe that language actually evolved as a means of internal expression,} Language, just like any other
technology produced by cultural evolution, has evolved over millennia through a repeated process of social transmission. Based on some recent estimates, human language evolved between 100,000 and 1,000,000 years ago (Barham & Everett, 2021; Dediu & Levinson, 2013; Perreault & Mathew, 2012), and has been adapting ever since, to suit the communicative needs of humans as the surrounding culture and social context also changed (Jackendoff, 2002). Of course, in truth this process of cultural evolution has resulted in the existence of many different languages. In fact, contemporary sources show that there are over 7,000 human languages in the world today (Eberhard et al., 2023). A number which does not include languages which existed in the past and of which we have no record, nor languages which currently exist but that have not been adequately documented.

The fact that there are so many uniquely different solutions to the same problem, the need to communicate, that have co-evolved across the human population is a fascinating phenomenon. It opens up discussions surrounding the idea of single or multiple points of origin of language (Chomsky, 2004; Nichols, 2011; Ruhlen, 1994). From an evolutionary perspective the question is if different human languages are the result of divergent evolution, where all modern languages share a common ancestor language that they diverged from, or convergent evolution, where language evolved independently in several groups in response to similar selective pressures (Pinker & Bloom, 1990). In fact, the existence of signed languages can be viewed as an argument in favour of convergent evolution, as many signed languages cannot be argued to have a direct phylogenetic predecessor in either other signed language or spoken languages. Independently of which of these evolutionary pathways is the correct one, the result of these evolutionary and cultural processes is a diverse collection of languages that vary along a number of dimensions, such as what syntactic structures they employ, what modality/modalities they use to express meaning, their phonemic inventories and vocabulary size (Evans & Levinson, 2009; Nichols, 1992). The fact that language can exhibit such a huge amount of diversity is a reflection of the fact that the hypothetical space of possible language structures is, theoretically, infinite (Berwick et al., 2011; Chomsky, 1965; Pinker, 1987; Wolff, 1982). Yet,

---

2 This is not to say that proto-language did not involve manual signals. In fact, some researchers argue that the first human language(s) were gestural (Corballis, 1999; Fay et al., 2022; Hewes, 1992; Tomasello, 2010). However, even if this is the case those early manual languages cannot be directly traced to modern signed languages.
in truth, human languages are not spread evenly across this hypothetical space, but rather cluster within it, such that many share specific features and strategies for communication (Culbertson, 2012; Kirby, 1999). The source of these similarities, sometimes called typological tendencies, is one of the key questions within linguistics. Are they simply the result of historical accidents? Are they driven by phylogenetic lineage within language history? Are they the result of common trends in language change, and possibly restricted by our physiological make-up?

The real answer is that they are probably the result of a combination of these causes, as research has found evidence of (at least some) influence of all these factors on certain linguistic phenomena. For example, certain word order patterns seem to be traceable in phylogeny, (Dunn et al., 2011; Hartung et al., 2022), and some research suggests that languages with certain phonological features, such as tone, are ecologically adaptive, with the result that we see higher concentrations of such languages in geographic locations with favourable environments for tone, such as warm areas with high humidity (Everett et al., 2015 although see Collins, 2017). Another proposed key player in influencing the shape of human language is the human mind. Our cognitive capacities, limitations, and biases are constantly active in our interactions with language, be it during speech/sign production or processing of auditory and visual linguistic information (Christiansen & Chater, 2008). Forming an understanding of the ways in which our cognitive system can contribute to generating these typological regularities provides us with a powerful tool to examine both present, past and future language contexts. Furthermore, the possibility that these biases may have a continuous, or iterative, effect on language structure as they apply in each new generation of language learners during the process of cultural transmission can inform our understanding of, not just the present patterns in language typology, but also the pathway of language evolution (Kirby et al., 2007; Thompson et al., 2016).

The research in this thesis explores a specific part of how our cognition might influence language structure, by focusing on the effects of cognitive biases. My aim is to examine under what conditions we can observe language-relevant cognitive biases in the behaviour of individuals, and how we can apply this knowledge to our understanding of how these biases apply as language evolves. In this work, I pay special attention to situations in which different types of cognitive biases come into conflict, pushing language structure in opposite directions. Examining such instances of competition allows us to not only isolate the individual contributions of cognitive biases, but also examine their relative strength under different conditions to see which type of bias is stronger in a given context. Due to this particular interest I focus on a subset of cognitive biases, namely those
which are active during language learning and language creation. The following sections will outline that we can observe interesting instances of competition and tension between such biases that can have profound effects on typological structure. There are a number of other types of cognitive biases that influence language structure, including those which are active during production (MacDonald, 2013), communication (Fedzechkina et al., 2023; Ferdinand et al., 2019), and processing (Gibson et al., 2019; Hawkins, 1983). Since several of these biases have already been examined in connection with learning biases (e.g. see work on the trade-off between compression and communication as structure-shaping factors in language evolution (Kirby et al., 2015)) I do not focus on them directly in this thesis.

With this specific interest in mind, I propose a new division of active cognitive biases, stratified according to what part of the linguistic system they influence, and what factors ground the existence of the biases. I motivate this distinction using evidence from studies of cognitive biases across phonological and syntactic phenomena. I then examine how different types of cognitive biases affect linguistic structure across both spoken and signed languages, based on typological and experimental evidence from both types of languages. In addition, I employ a set of experimental tools that allow me to generate linguistic contexts that draw upon different cognitive mechanisms, as well as compare these experimental results to typological data. Finally, I make predictions concerning the contexts in which we might expect to see each type of bias guiding linguistic behaviour, by organising tasks and contexts into a structured scale based on the amount of linguistic innovation that is needed in each individual task (the Scale of Innovation).

1.2 Chapter outline

The remainder of this chapter will be dedicated to providing the relevant theoretical, typological and experimental background to the research presented in this thesis. I will define key concepts, provide novel definitions and conceptualisations of some concepts in light of the body of evidence that is currently available, and situate the experimental work that makes up this thesis among similar and relevant research across linguistics, cognitive science and psychology. Section 1.3 outlines the concept of typological regularities, as well as proposed explanations for these patterns, and gives examples of such regularities across spoken and signed languages, with a special focus on word order phenomena. Section 1.4 gives a working definition of the type of cognitive biases which will be
the focus of the thesis, then illustrates the effects of complementary and competing cognitive biases on typological structure. I also introduce a new two-tiered division among cognitive biases based on their sphere of influence on language structure and the motivation for why they exist. I illustrate these insights with examples of phenomena from syntax, phonology, and beyond. Section 1.5 focuses on describing a proposed Scale of Innovation, a theoretical construct which places different experimental contexts along a scale based on the amount of innovation that the task/context requires. With this scale established I then map the effects of certain types of cognitive biases onto different parts of this scale. Based on this background, Section 1.6 outlines the typological phenomena I use as a case study in this thesis to examine the proposed bias division, the effect of bias competition, and the way that these biases interact with the Scale of Innovation. Finally, in Section 1.7, I give a road-map for the content of the thesis by establishing the main research questions and how attempting to answer those questions will contribute to our wider understanding of how cognitive biases influence typological structure across linguistic domains, and the role they play during language evolution.

1.3 Typological regularities

Just as cumulative cultural evolution gives rise to many different versions of the same technology, the distribution of languages across the world can be seen as the resulting distribution over the cultural variants of different languages after a continuous period of cultural transmission. The focus of the field of linguistic typology has long been to try and map the features of this distribution. This is achieved by documenting similarities and differences between languages, and identifying patterns in the way these features group across the world (Bickel, 2007; Comrie, 1988). Early typological work mapping these patterns, such as that conducted by Greenberg (1963), was fairly limited in terms of the number of languages it could reasonably use as the basis for these generalisations. This was due to a number of factors, mainly relating to accessibility of detailed, high-quality linguistic descriptions of a sufficient number of the world’s languages, combined with the time it took to consolidate and compare these sources manually. Due to this, many early typological studies were based on a small subset of languages, where samples often consisted of closely related languages and the categories and features under investigation were those formulated by a Eurocentric scientific tradition (Scalise et al., 2009). Over-sampling from a specific language family when
trying to draw typological generalisations can skew the study results, resulting in inaccurate estimates for how common or rare a certain language feature is (Dunn et al., 2011). For example, in Greenberg’s early work he proposed the existence of forty-five typological regularities. These proposals were based on a sample of only thirty languages which, although he worked hard to balance across language families, he himself recognised as being largely based on “convenience” and personal experience with the selected languages (Greenberg, 1963, p. 41).

In recent years, several different databases of typological data have been collated and published for public and academic use. These include earlier sources, such as Ethnologue (Eberhard et al., 2023/1951), the more recent World Atlas of Language Structures (Dryer and Haspelmath, 2013/2008) and Glottolog (Hammarström et al., 2022/2011), and most recently Grambank (Skirgård et al., 2023). These provide invaluable resources to current linguists who are interested in studying typological regularities, as they allow us to draw comparisons for specific features across a large number of languages from a wide selection of language families. Two of the more recent databases, WALS and Grambank, contain data from 2,662 and 2,467 languages respectively and span over 200 language families and many isolates. The more recent source, Grambank, features a very high number of languages for each language feature that it covers, often including classifications of over 2,000 languages for each feature, whereas WALS tends to have lower language counts (sometimes only a few hundred) for each individual feature. Combined, these sources are the most comprehensive resources on linguistic typological that exist to date.

Something worth noting about the history of typological research is that, in many of the early studies, signed languages were not considered for inclusion in these investigations (Comrie, 1989; Greenberg, 1963). In some cases researchers make explicit mention of why they do not choose to include signed languages in their samples, citing motivations such as the difficulty in evaluating whether the feature they are studying is present in a comparable way in signed languages (Leeson & Saeed, 2012; Zeshan & Palfreyman, 2017). More recent editions of the available typological resources do contain data from signed languages, either in their default sample or in specific dedicated chapters and data sets (Eberhard et al., 2023; Hammarström et al., 2022; Skirgård et al., 2023; Zeshan, 2013). As belated awareness of the value that studying signed languages bring to the scientific community has increased, more and more typological work has been conducted that specifically focus on features of signed languages (Aronoff et al., 2004; Ferrara et al., 2022; Zeshan & Palfreyman, 2017). These studies examine the existence of parallel patterns in signed and spoken languages, such as ordering
of linguistic elements (Coons, 2022; Napoli & Sutton-Spence, 2014), as well as features which are unique to signed languages, such as the use of simultaneity (Morgan, 2002; Napoli & Sutton-Spence, 2010; Slonimska et al., 2020), pronominal pointing (Cormier et al., 2013), and constructed action (Cormier et al., 2015). This work is invaluable to linguists who are interested in the cross-modality of typological tendencies, as having data from both spoken and signed languages give us a richer understanding of the modality specificity and generality of any observed patterns (Zeshan & Palfreyman, 2017).

1.3.1 Language universals

In early typological research, and much subsequent discussion surrounding the topic of linguistic diversity, the concept of language universals have been a central notion (Christiansen et al., 2009; Comrie, 1989; Greenberg, 1963; Kiparsky, 2008; Plank, 2011). Greenberg’s original forty-five typological patterns were proposed as language universals, either taking the form of absolute universals (see example (1)), which were expected to be exceptionless across all the world’s languages, or they were formulated as statistical universals like Universal 4 (see example (2)) that were said to occur more or less often than would be expected by chance (Greenberg, 1963). These types of typological patterns can be found across many linguistic domains, although Greenberg’s main focus was on syntactic and morphological phenomena. Subsequent work has also argued for the existence of similar organising principles in phonology and phonetics (Blevins, 2004; Kiparsky, 2008) and semantics (Goddard, 2001; Wierzbicka, 1996).

(1) Universal 3

“Languages with dominant VSO order are always prepositional.”

(Greenberg, 1963, p. 45)

(2) Universal 4

“With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional.”

(Greenberg, 1963, p. 45)

The idea that some of these language universals should be exceptionless is, arguably, predicted by certain approaches to generative grammar as a side-effect of the way this theoretical approach claims that Universal Grammar (UG) provides a solution to Plato’s Problem. In linguistics this problem relates to hypothesis
selection during language acquisition. How do children choose the correct grammar for their target language when an infinite number of possible grammars are compatible with the data they encounter? While the space of theoretically possible grammars is infinite, real human languages seem to occupy bounded regions within this infinite space of possible languages and languages tend to cluster together within these regions based on shared structures and features. So the question is, how do children acquiring language home in on the space that is occupied by actual human languages? According to traditional conceptualisations of UG, this is achieved through the existence of constraints on possible linguistic structure which are both innate and inviolable (Chomsky, 1965, 1966). These constraints serve to delimit the set of possible hypotheses that a language learner needs to entertain as possible target grammars during acquisition by ruling out those grammars which violate the innate constraints (Baker, 2001; Newmeyer, 2005). As a result, linguistic systems which are theoretically possible, but ruled out by the constraints in UG, are predicted to never occur among the world’s languages and therefore suggests the existence of absolute universals of the type that Greenberg formulated for Universal 3 (example 1).

Despite this prediction, finding true exceptionless universals has proven difficult. As detailed analyses of the grammatical structure of an increasing number of diverse languages become available, more accurate estimations of the reliability of universals are available (Bickel, 2007). In fact, proving that a universal is truly absolute is practically impossible, no matter the size of the sample available to researchers (Piantadosi & Gibson, 2014). Over time, and with more available data, universals which were previously thought to have no exceptions are found to only be statistical tendencies. For example, current data from WALS shows that Universal 3 has 6 instances of languages with VSO order and postpositions (Dryer, 2013b, 2013e), the direct opposite of the structure proposed by the universal. These languages include Guajajara, a language spoken in Brazil, and Majang, a language spoken in Ethiopia. Similarly, Universal 18, which stated that languages with prenominal adjectives also had prenominal numerals, had no exceptions in Greenberg’s original sample (Greenberg, 1963, p. 51). However, WALS now shows that although there are 251 languages that follow this pattern, there are 37 languages which have the opposite order (Dryer, 2013a, 2013d). Given these discrepancies between posited absolute universals and modern data, more recent discussions of language universals tend to formulate them as statistical tendencies. Notably those tendencies still mostly align with the directionality of previously stated absolute universals (Christiansen & Chater, 2008; Dryer, 1992, 1998; Evans & Levinson, 2009). In line with this changing view of language
universals my default throughout this thesis will be to use the terms typological regularity, typological tendency, or typological pattern to refer to these phenomena. I believe these names convey the probabilistic nature of these typological regularities. I also think that this terminology is more appropriate given the nature of what I will argue contributes to the existence of these patterns, and also acknowledges the constant discoveries of new dimensions of diversity across languages which make our estimates of typological patterns ever-changing.

1.3.2 Explanations of typological regularities

If we follow the reasoning that the existence of absolute universals is hard to prove, and reject the idea that these typological tendencies are the result of the type of hard constraints or parameters which generative syntax proposes constitute our language faculty, then we are still left with the need to explain why we observe these typological tendencies. If we are not hard-wired to disallow certain language structures then why are languages not spread more evenly across the hypothetical space of possible languages? There are a large number of proposed explanations for these statistical tendencies which are popular in different sub-fields of linguistics. One approach is to appeal to the existence of common pathways during language change, arguing that these shared processes result in typological tendencies (Aristar, 1991; Bybee, 2008; Collins, 2019; Ohala, 1992). One phenomenon which has seen successful application of this idea is studies of grammaticalisation, which is the process where language units go from fulfilling a lexical function to fulfilling a grammatical function, or where a unit that is already grammatical assumes an even more grammatical function (Heine et al., 1991). This type of definition of grammaticalisation assumes that this process is unidirectional, such that you do not tend to see the opposite, degrammaticalisation pattern, where language units that fulfil highly grammatical roles change to fulfil lexical roles instead (Lehmann, 1995).

The unidirectionality of this process has been argued to constitute a universal property of language in and of itself, serving as a useful tool in language reconstruction (Haspelmath, 2004) and which in turn can give rise to the type of typological patterns discussed previously (Norde, 2009). A prototypical example of the way grammaticalisation can be used to explain a typological pattern is the role it plays in generating inflectional systems. A common grammaticalisation process observed across a number of languages is that adpositions or pronouns tend to fuse with verbs to create inflectional systems, but it is very rare to find the opposite process, where inflectional affixes detach from verbs to form new pro-
nouns or adpositions (Kiparsky, 2011). Similarly, adpositions tend to originate in grammaticalisation from verbs in verb phrases, meaning that the typological tendency for adpositional placement (prepositions or postpositions) to correlate with the order of verb and object in verb phrases could originate in this historical process (Bybee, 1988; Dryer, 2019; Plank, 2011). However, more and more evidence is being presented which challenges the unidirectional of grammaticalisation, suggesting that not all typological tendencies are the result of fixed/universal pathways in language change (Haspelmath, 2004; Norde, 2009).

Another explanation for the existence of typological regularities is that they may be the result of phylogenetic history. Some language features may simply have arisen in the proto forms of many of the present-day languages that went on to have widespread and diverse linguistic lineages. This explanation is therefore not so much about convergence of language change, but about consistency of features within a language family across diachrony (Bickel, 2007). This type of theory has promoted several attempts at examining how many of the typological regularities could be explained by language phylogeny (Dunn et al., 2011; Hartung et al., 2022; Piantadosi & Gibson, 2014). To do this researchers have built computational phylogenetic models that take contemporary theories about language families into account when examining if word order patterns, like those formulated by Greengberg, can be considered lineage-independent. Several of these models have found that many of these word order correlations are strongly lineage-specific, and therefore cannot be considered as general typological tendencies (Dunn et al., 2011; Hartung et al., 2022), although the extent to which this is true and how many of the typological tendencies are found to have some lineage-independence varies across studies (Jäger & Wahle, 2021).

Some theories rely more on the interplay between human and natural-world physiology to explain cross-linguistic tendencies. These explanations of language features appeal to how our physiological limitations and the structure of our audio/visual production and processing systems function. Due to this, they are usually found in studies of phonological typological tendencies, as they provide some grounding for these patterns in our abilities to perceive contrasts between sounds/gestures and, in turn, our abilities to produce those sounds and gestures. For example, in a study examining the correlation between aridity and the existence of tonal contrasts across the world’s languages Everett et al. (2015) found that tone is more likely to be a contrasting language feature in locations where the local climate is humid and hot. The physiological motivation behind this is that tone requires precise control of the vibration of the vocal folds, something which is harder to achieve in colder and more arid climates. This work, and
others of a similar nature (Maddieson et al., 2011; Munroe et al., 2009), suggest that language structure may be, to some extent, physiologically adaptive to its environment.

These explanations may all be contributing factors for the existence of typological tendencies, and might have more or less influence depending on the specific linguistic domain in question. However, in light of the shifting view of universals as statistical tendencies, another research field has emerged which considers that, just like the typological patterns themselves are probabilistic, so are the constraints which give rise to these patterns. This approach, which still views typological patterns as a meaningful reflection of human cognition, rather than accidents of history, argues that typological tendencies are the result of cognitive biases which favour or disfavour certain language structures (Culbertson, 2012; Culbertson & Kirby, 2016; Culbertson & Smolensky, 2012; Culbertson et al., 2013; Hudson Kam & Newport, 2009; Smith & Wonnacott, 2010; Wilson, 2006). Some proponents of this kind of approach claim that these biases are, essentially, external to the language system and constitute general cognitive principles which may have specific effects within the system of language (Chater & Vitányi, 2003; Culbertson & Kirby, 2016; Hupp et al., 2009). Examples of such cognitive general biases may include memory constraints, or biases involved in general information processing. Connecting this approach back to earlier definitions of what constitutes the language faculty, cognitive biases could be said to fall into what Chomsky calls the “third factors”. The first two factors are genetic endowment and experience, whereas the third factors are formulated as “principles not specific to the faculty of language” but which still influence the shape of language (Chomsky, 2005, p. 6). These more general cognitive biases could be seen as such principles, as they guide our behaviour and our interaction with language without being specifically dedicated to the language faculty.

Other researchers who pursue the idea that cognitive biases have an impact on language argue for there being a slightly closer relationship between cognitive biases and the language faculty, such that some biases are specific to the way we interact with language but do not necessarily influence our behaviour much beyond language (Culbertson & Adger, 2014; Culbertson et al., 2012; Martin et al., 2020; Schouwstra & de Swart, 2014). Independently of exactly where these biases are localised and their generality, they are fundamentally different from the inviolable constraints that previous generative approaches promoted, as they are violable and so, under this approach no linguistic system is deemed as impossible. Instead, this approach argues that those grammatical systems which satisfy our biases end up being over-represented across typology as they are, for
example, easier to learn (Christiansen, 2000; Culbertson, 2012; Culbertson et al.,
2012; Wilson, 2006) or more likely to be created as the default language structure
(Goldin-Meadow et al., 2008; Marno et al., 2015; Motamedi, Wolters, Naegeli,
et al., 2022). On the flip-side, those systems which do not adhere to the biases
are disfavoured as they are harder to learn or less likely to be created, leading
to those systems being rare across typology (Culbertson, 2012). As a result,
these biases restrict the infinite number of possible grammars in a similar way
to the hard constraints of UG, but in a probabilistic rather than absolute sense.
This approach provides a different solution to the hard problem of induction
that children are faced with during acquisition. Instead of some grammars being
ruled out due to UG, and therefore never considered as possible hypotheses during
acquisition, cognitive biases mean that learners search the space of grammars that
adhere to their biases first, before considering other alternatives. This makes some
grammars which are compatible with the data more likely to be considered early,
and selected as the target grammar by learners (Chater & Christiansen, 2010).
These biases can be conceptualised as acting in a similar way to Bayesian priors
over hypotheses, such that they increase the probability of selecting favoured
hypotheses over other alternative hypotheses (Culbertson et al., 2013; Reali &
Griffiths, 2009). In this way, cognitive biases can both guide acquisition, and over
time make areas within the space of possible languages more densely populated
with real human languages. It follows from this that cognitive biases of this type
would also give rise to typological regularities, but that these are predicted to
be probabilistic rather than absolute (Baronchelli et al., 2015; Culbertson et al.,
2013).

1.4 Cognitive biases and linguistic behaviour

This next section will provide some additional detail about how I specifically em-
ploy the term cognitive bias in this thesis. The word bias is not only used within
the context of language, but rather applies to a wide range of concepts and phe-
omena in the world. We might talk about social biases relating to prejudices
against certain groups and individuals, or refer to the mechanistic biases that are
present in machine learning models due to their design or skews present in their
training data. When I employ the term cognitive bias in this thesis I use it to
signify a specific concept within cognitive science and linguistics, namely features
of the human cognitive system that might make us favour or disfavour certain
types of patterns and/or structures (Baronchelli et al., 2015; Culbertson, 2012,
This definition of cognitive bias involves both more general biases that impact our wider cognitive abilities, such as memory constraints (Hudson Kam & Chang, 2009) and our general expectation that random events should be uniformly distributed, as well as potential biases that are specific to our interaction with linguistic tasks and data.

The identification of these biases, and the study of their impact on human behaviour, is of interest to all those who seek to understand the way the human cognitive system gives rise to human behaviour. For example, researchers studying machine learning and natural language processing might implement approximations of these general restrictions on human cognition in hopes of making their models generate more human-like behaviour. Cognitive biases like memory restrictions have been successfully implemented in a number of deep learning contexts, showing that such biases give rise to better generalisation between training data and unseen testing items as it limits the model’s ability to simply encode large amounts of the language structure and forces it to construct more robust generalisation strategies (Resnick et al., 2019), giving rise to more regular languages (Conklin & Smith, 2023). Similar strategies were previously employed in other modelling frameworks, such as iterated learning models, where the size of the data bottleneck (i.e. how much data is passed from one generation of artificial agents to the next generation) is shown to impact the emergence of both compositionality and regularity in communicative systems (Brighton et al., 2005; Kirby, 2001).

One of the key methodological tools available to researchers who wish to study how cognitive biases shape human languages in a behavioural setting is the use of artificial language learning (ALL) experiments (Culbertson, 2023; Fedzechkina et al., 2016). Broadly, this class of behavioural experiments involves training participants on miniature made-up languages that have certain properties of interest to the researchers. By building these experiments, researchers can examine things like learnability differences across different language features, and the inferences people make in the face of limited linguistic input in a controlled environment.

This method has been employed across a number of different sub-fields of linguistics, including syntax, morphology and phonology. One of the earliest research fields to adopt this type of method is the study of statistical learning. Statistical learning experiments examine participants’ abilities to extract distributional information that allow them to form representations of things like word boundaries and grammatical rules/categories from a stream of audio input (Saffran et al., 1996). Crucially, these types of ALL studies do not involve any semantic component. Other studies, examining things like compositionality (Kirby
et al., 2008, 2015) or cues for noun class membership (Culbertson et al., 2019) have meanings paired with the words in the artificial language. The size of these artificial languages can vary quite widely from experiment to experiment, with some including as many as 51 different words and requiring several days worth of training (Hudson Kam & Newport, 2009), to others with only two meanings (Motamedi, Wolters, Naegeli, et al., 2022). The languages themselves also vary in terms of how artificial they are. Some use completely novel lexical items (e.g. Martin et al., 2020), others employ semi-artificial languages by having the lexical items be words in participants’ native language but in an artificial grammatical structure (Culbertson & Adger, 2014; Meisezahl et al., 2023), or by capitalising on signal-sign correspondence (e.g using the word buzzo to denote the meaning of ‘bee’; Keogh et al. (2023), or the word twit to denote the meaning of ‘bird’; Atkinson et al. (2018)). These latter options alleviate some of the burden of word learning, allowing researchers to build more complex artificial languages that are still learnable in a short period of time.

Artificial language learning experiments allow researchers to probe cognitive biases related to specific language features in an interesting and precise way by measuring, for example, differences in learning rate and grammaticality judgements between different language types and conditions. However, participants’ experience of their own native languages can transfer to their interaction with the artificial languages they are exposed to in the experiments. One way in which this type of transfer can be limited is to create gesture-based artificial languages instead of spoken artificial languages. The transfer-limiting effect of changing communicative modality was first discovered in experimental studies that did not involve learning an artificial language. Instead, in these studies hearing participants with no proficiency in any sign language had to improvise new linguistic expressions for specific meanings using only their hands. The resulting structures that these participants generated were remarkably similar, and showed little influence of their individual native spoken language structure (Goldin-Meadow et al., 2008). This methodology has been successfully employed to study biases relating to phenomena such as word order in the noun phrase (Culbertson, Schouwstra, & Kirby, 2020; Jaffan et al., 2020) and basic word order preferences (Futrell et al., 2015; Gibson et al., 2013; M. L. Hall et al., 2013, 2014). More recently it has also been used in combination with general artificial language learning methodologies to study the evolution of linguistic structure (Motamedi-Mousavi, 2017).

By examining the results of a number of studies that use methods similar to those outlined above I will illustrate how the current evidence supports a two-tiered division between cognitive biases that are either system-wide, or category-
specific. I base this division on what motivates, or grounds, these biases and how there is a clear difference in the source of this grounding between category-specific and system-wide biases. Furthermore, the two types of biases seem to affect people’s interactions with different aspects of the linguistic data, either influencing the systematic structure within the wider language system, or biasing the way specific categories of items are expressed or ordered linguistically. Examples and evidence in the following section are drawn from studies in both syntax and phonology, suggesting that this division goes beyond just syntactic biases, despite the fact that the experiments in this thesis mainly focus on word order phenomena. Furthermore, I refer to evidence from studies in fields other than linguistics where such evidence can inform our understanding of linguistic phenomena.

1.4.1 Simplicity phenomena

An important set of phenomena which has been ascribed to the influence of cognitive biases are those connected to the concept of simplicity. Appeals to a preference for simplicity extend far beyond interactions with linguistic data, but one of the main ways in which simplicity has been argued to influence linguistic systems is during tasks that require inductive reasoning, such as language acquisition tasks (Chater & Vitányi, 2003). A bias for simplicity has been suggested as a guiding principle that limits uncertainty by influencing language learning behaviour. It predisposes learners to acquire simpler grammatical systems and rules that they can then easily generalise across new forms in the language (Culbertson & Kirby, 2016).

Structural complexity in the syntactic domain can be defined as the number of syntactic rules that are needed in a grammar to capture the patterns and variations in the language (Culbertson, Franck, et al., 2020). Expressed in an information theoretic way, complexity can be measured as the extent to which the data in the language is compressible, i.e. how many bits are needed to encode the language structure (Teal & Taylor, 2000; Wolff, 1982). Languages that exhibit systematic syntactic structure allow for more compression (fewer bits to encode), as these systems rely on higher level rules that capture grammatical patterns across many parts of the language. The grammars of these languages can therefore be thought of as simpler than languages with extensive arbitrary structure (e.g. languages with extensive unconditioned variation) since such languages require a larger number of rules or lexically encoded exceptions (Culbertson & Kirby, 2016; Kirby et al., 2008; Smith & Wonnacott, 2010). Simplicity has thus been
proposed as a bias which predisposes humans to favour compressible syntactic systems.

There are a number of properties that can make a grammar less complex and more compressible. For example, a language that exhibits consistent ordering between heads and dependents needs fewer rules to describe the linearisation of phrasal elements compared to a language that has different ordering rules for different types of dependents (Culbertson & Kirby, 2016). Languages with consistent head directionality are referred to as harmonic languages, and such grammars can posit a single high-level ordering rule which can then be generalised to all head-dependent pairs across the system. On the other hand, languages with inconsistent orders between different heads and dependents need to posit more rules to capture the linearisation of individual phrase and possibly individual dependents (Culbertson, Franck, et al., 2020, see Table 1.1 for illustrative examples of harmonic and non-harmonic language systems). In fact, harmonic structure is widespread in typology. Greenberg (1963) noted that, for example, languages with Object-Verb order tend to be postpositional (i.e. head-final) whereas languages with Verb-Object order tend to be prepositional (i.e. head-initial). Dryer (1992) presented further data showing that there are in fact 24 pairs of elements that correlate with the order of the verb and object.

<table>
<thead>
<tr>
<th>Harmonic</th>
<th>Non-harmonic</th>
</tr>
</thead>
<tbody>
<tr>
<td>XP $\rightarrow$ X ...</td>
<td>NP $\rightarrow$ N...</td>
</tr>
<tr>
<td>VP $\rightarrow$ ...V</td>
<td>PP $\rightarrow$ P...</td>
</tr>
</tbody>
</table>

Table 1.1: Illustrative example of how a harmonic language requires fewer linearisation rules than a non-harmonic language.

Due to the compression benefits of harmony, a wide variety of research has explored the possibility that there is a cognitive bias in favour of harmonic languages, and specifically pursuing the idea that harmonic structure imparts a learning advantage over other language systems. Among these studies Christiansen (2000) found experimental evidence for a preference in the direction of harmonic patterns. In his study, participants trained on a harmonic language were better at identifying ungrammatical phrases at test, compared to participants trained on a non-harmonic language, suggesting that the language with a harmonic structure was easier to learn. Combining the observation that learners regularise variation and the hypothesis that learners show a preference for harmonic word orders, Culbertson et al. (2012) predicted that learners exposed to variable noun phrase word order would regularise harmonic NP orders to a greater
extent than non-harmonic orders. This prediction was borne out in their study with adult participants, and similar preferences for harmonic orders were later found with child learners as well (Culbertson & Newport, 2015, 2017). Goldberg (2013) argued that the findings in Culbertson et al. (2012) could be a reflection of abstract transfer of a harmonic preference based on participants’ L1 experience (participants were English speakers, which has a harmonic NP order), rather than evidence for a general bias for harmonic orders. Yet further follow-up studies with speakers of languages with non-harmonic noun phrase orders (Hebrew and French) established that the results were not merely a reflection of participants’ native language structure (Braquet & Culbertson, 2017; Culbertson, Franck, et al., 2020), but also extended to populations with different native language experience.

Additional evidence supporting the idea that language structure is affected by biases that favour systematic organisation of linguistic elements includes computational simulations and experiments that employ iterated learning designs. This methodology involves training individuals (artificial agents in this instance) on a language system and then having them produce items from that language. This output is then used as the training input for the next generation of learners, forming a transmission chain. Simulation work in this area has found that compositional structure can emerge from holistic languages, resulting in a simplification of the language grammar. Holistic languages are those where complex meanings are denoted by holistic signals, such that no sub-part of the signal refers to any sub-part of the meaning. This makes holistic languages incompressible and the system needed to encode them is highly complex. The most efficient encoding of a holistic language is a list of every signal and its corresponding meaning in the whole language. Compositional languages, on the other hand, allow for compression in the grammar encoding as signals are made up combinations of individually meaningful sub-parts (Kirby et al., 2015). Simulation studies examining the emergence of compositional structure observed that the size of the bottleneck (i.e. how much data learners were exposed to) is crucial to this process of simplification due to the fact that limited amounts of data forces learners to make generalisations of the type that favour compositionality (Brighton, 2002; Brighton et al., 2005; Kirby, 2001; Smith et al., 2003; Zuidema, 2002). Furthermore, in work by Kirby et al. (2007) it was highlighted that the process of iterated learning itself could magnify a weak prior bias for simple structures, meaning that strong biases were not necessary for population wide effects of simplification to occur. These results were also mirrored in experiments with human participants, where similar processes of regularisation and evolution of compositional structure were observed.
over generations of learners (Kirby et al., 2008, 2015; Reali & Griffiths, 2009).

Other processes of simplification have also been observed in communicative tasks that do not involve language. For example Garrod et al. (2010) found that drawings, used to communicate different meanings, were drastically simplified when transmission was horizontal (i.e. repeated interaction between individuals), but that iterated learning across a chain of individuals did not result in the same kind of simplification. Whereas Theisen-White et al. (2011) found that, when a communicative element was added to the iterated learning design, the drawings showed signs of simplification through the evolution of compositional structure.

A key factor when considering the results of these non-linguistic tasks in connection with the work on linguistic data is the distinction between simplicity of individual items, versus simplicity of systems. Compositional structure is only simple if considered at a system level, as the combination of individually meaningful elements into larger more complex meaningful elements allows the grammar of that language to be simple, but can make the internal structure of items (e.g. words and phrases) more complex than holistic representations of meaning which are not the result of combinatorial operations. Thus the study by Garrod et al. (2010) observed simplification at the item level as a result of horizontal transmission, whereas Theisen-White et al. (2011) observed simplification at the system level through the emergence of compositional structure when a communicative element was added to an iterated learning task.

A related set of results are those from research that examine people’s tendency to reduce unpredictable variation in linguistic data. If this tendency is truly caused by a bias for simplicity is currently an area of active debate, with some researchers arguing that unpredictable variation is reduced independently of such a preference (Ferdinand et al., 2019), whereas others have argued that it is a reflection of the wider simplicity bias that is active during learning (Culbertson & Kirby, 2016). With that said, extensive experimental work has been conducted that investigates how the processes of language learning and production affect levels of variation in language. Hudson Kam and Newport (2005, 2009) trained participants on an artificial language with inconsistent use of different determiners and observed that, at test, participants regularised the variation of determiners to establish a majority form. They also identified that adults, although exhibiting regularisation behaviour in conditions with high variance (i.e. when there was a high number of different determiners) would reproduce more of the input variation than children would. Instead, children seem more likely to extend the use of majority orders, compared to adults, suggesting that the bias for simplicity might be especially strong in children and during early acquisition
(Newport, 2016). In another artificial language learning experiment, Smith and Wonnacott (2010) found that the process of iterated learning, where the output of one generation of learners becomes the input to the next generation, could also lead to the reduction of unpredictable variability. Participants exposed to a language with random use of different plural markers evolved a grammar over time where the variation was conditioned on noun type. Similarly, Reali and Griffiths (2009) trained participants on a language that had two possible words to express each meaning in the language and, during training, the probability with which each word occurred varied. They found that, although individual learners would generally closely match the levels of variation present in their input, over generations the language shifted to become more regular. The same type of regularisation behaviour has also been observed in situations of natural language learning. For example, there is a detailed case study of a deaf child who acquired American Sign Language (ASL) from hearing parents who were late learners of ASL. In this context the parents’ language productions included a lot of unpredictable morphological variation, yet the child regularised this variability in his own productions (Singleton & Newport, 2004). Furthermore, patterns of variation reduction have also been observed in non-linguistic learning tasks, although the regularisation tendency seems to be stronger in tasks with linguistic stimuli (Chater & Vitányi, 2003; Ferdinand et al., 2019). Even if these results are not a reflection of a simplicity bias, they are a reflection of individuals’ preference for systematicity across language structure, and could help explain why unconditioned, free variation is comparatively rare in language (Givón & Slobin, 1985).

Syntactic phenomena are not the only source of evidence for the effect that a bias for simplicity can have on typological structure. Similar explanations can also be presented for a variety of patterns observed in phonological typology. A tendency for phonological systems across the world’s languages to exhibit structural simplicity has been observed in typological data (Clements, 2003), especially when considering the number of contrastive features languages have in their inventory, or in the complexity of the phonological alternations in their grammar. Drawing on insights from research in psychology looking at categorisation, participants tend to show better and faster learning of patterns and categories that require them to encode fewer stimuli features (e.g. shape, colour, pattern etc.), illustrating a bias towards simpler patterns (Feldman, 2006; Nosofsky et al., 1994; Shepard et al., 1961), which could also apply to the encoding of phonological features (e.g. height, rounding, length etc.) during phonological learning. This parallel is evident in the design of many experimental investigations of the role
that simplicity plays in phonological learning, since participants are often exposed to data designed to teach them phonological rules that require different numbers of features to be encoded by the learner (for a review, see Moreton & Pater, 2012a).

Just like in the case of word order harmony, the possibility of there being a cognitive bias in favour of phonological simplicity has been examined using a wide variety of artificial language learning tasks. In studies that compare tasks requiring participants to attend to a single feature and tasks where participants have to attend to several features to distinguish the class of sounds involved in a phonological alternation, results consistently show that the single feature rule is easier to learn (Conrad et al., 1977; Pycha et al., 2003; Saffran & Thiessen, 2003; Skoruppa & Peperkamp, 2011). Similarly, studies that compare the number of feature changes mandated by an alternation also find that alternations involving a single feature change (see Figure 1.1a) were acquired better than alternations involving several feature changes (see Figure 1.1b) (Peperkamp et al., 2006; Skoruppa et al., 2011; White, 2014). In addition, a more recent study by Durvasula and Liter (2020) found that, although learners can make use of multiple rules to account for an alternation, they may only ever encode the simplest ones. To elaborate, if an alternation can be accounted for by two simple rules that each refer to a single feature then learners will acquire those rules, but not the more complex rule that combines the two features.

\[
\begin{align*}
t \rightarrow d &= \left[ \begin{array}{c} -\text{continuant} \\ -\text{voice} \\ +\text{coronal} \end{array} \right] \rightarrow \left[ \begin{array}{c} -\text{continuant} \\ +\text{voice} \\ +\text{coronal} \end{array} \right] \\
\text{(a) Alternation with one feature change.} \\
&t \rightarrow \delta = \left[ \begin{array}{c} -\text{continuant} \\ -\text{voice} \\ +\text{coronal} \end{array} \right] \rightarrow \left[ \begin{array}{c} +\text{continuant} \\ +\text{voice} \\ +\text{coronal} \end{array} \right] \\
\text{(b) Alternation with two feature changes.}
\end{align*}
\]

Figure 1.1: Examples of alternations that involve either one or two feature changes.

A further set of experiments have investigated how the complexity of relations between phonological features affects learning and generalisation behaviour. For example, dependencies between instances of the same feature are typologically more common than dependencies between different features (Archangeli & Pulleyblank, 1994; Rose & Walker, 2004). This pattern has been found in several
experiments (Lin, 2009; Moreton, 2008, 2012; Wilson, 2003), all of which display better learning and higher familiarity effects when two instances of the same feature are involved in the stimulus alternation. However, in some experiments this difference was not observed (e.g. Kuo, 2009; Seidl & Buckley, 2005).

Taken together, the evidence from both syntax and phonology supports the existence of biases that relate to the quantity of rules or features needed to define a syntactic system or phonological alternation. These biases seem to influence individuals, especially during experimental tasks that involve learning of a language system, making those systems/rules which are favoured by these biases easier to learn. Given that these biases seem to be active in individual language learners this means that these biases have the potential to apply iteratively, with each new generation of language learners, resulting in cumulatively larger effects on language structure over time (Reali & Griffiths, 2009; Thompson et al., 2016). This cumulative effect could be the reason that we see typological patterns that seem to reflect these biases, such as the typological tendency for harmonic word order and phonological feature economy.

1.4.2 Naturalness phenomena

Another category of typological regularities in phonology relates to the concept of phonological naturalness and includes studies of the prevalence of natural phonological rules (e.g. Garrett & Johnson, 2013; Hooper, 1976; Rose & Walker, 2011). Yet what characterises a rule as “natural”, and what role such a feature plays in influencing actual typological patterns varies between different theoretical approaches to phonology (Dinnsen, 1980). One type of approach argues for the idea that what counts as natural in phonology is determined by phonetic principles (Archangeli & Pulleyblank, 1994; Blevins, 2004; Donegan & Stampe, 1979, 2009; Stampe, 1979). According to these theories, natural phonological rules are rules that align with general phonetic pressures and processes. Among these theories there are disagreements regarding the level at which naturalness affects typology. Some argue that naturalness only contributes to typology through diachronic processes, but does not involve active synchronic biases in the minds of individuals (Blevins, 2004; Hellberg, 1978; Moreton, 2008; Ohala, 1993). For example, Ohala (1993) claims that instances of hypo- and hyper-correction can lead to the typological dominance of natural rules due to general phonetic pressures that then grammaticalise to phonological sound changes (see Mailhot, 2013, for a simulation account of this type of process with regards to vowel harmony). Theories that do support the existence of phonetically motivated phonological
biases claim that such biases have synchronic effects on typology, by influencing the behaviour of individual learners (Archangeli & Pulleyblank, 1994; Donegan & Stampe, 1979; Hayes et al., 2004; Hooper, 1976). According to these theories, a phonological bias for naturalness would predispose learners to acquire phonetically motivated, natural rules more easily than unnatural ones since these align with their existing phonetic knowledge. Better learning would then lead to better population-wide retention of these rules compared to unnatural ones. This would eventually ensure a typological skew in the direction of natural rules (Martin & Peperkamp, 2020; Wilson, 2006).

Identifying experimental evidence that reliably corroborates the synchronic effect of naturalness on phonological behaviour has proven relatively difficult (see Moreton & Pater, 2012b, for a review). In some cases this lack of conclusive evidence can be ascribed to confounding factors in the experimental design, whereby naturalness effects become obscured by effects related to simplicity or salience. For example, Schane et al. (1974) found that participants learned a rule of word-final consonant deletion in a cluster better than deletion of a word final consonant in an intervocalic context. These results would support the naturalness hypothesis but, as Moreton and Pater (2012a) point out, participants could have learned rules where they simply counted the number of consonants allowed between vowels, resulting in the unnatural rule being more complex and therefore the difference in learning could be ascribed to a simplicity rather than a naturalness bias. Similarly, Gerken and Bollt (2008) conducted experiments with infants to investigate the naturalness motivation for the typological dominance of coda-conditioned stress patterns. They did this by comparing coda-conditioned stress patterns to onset-conditioned stress patterns and found that the former was learned better. However, the coda-conditioned pattern depended on the absence or presence of a consonant, whereas the second depended on features of an ever-present consonant, making the complexity of the two patterns hard to equate.

The effect of this type of conflation between naturalness and simplicity can also go in the opposite direction, obscuring the potential presence of a naturalness effect. For example, Glewwe (2017) compared extension of voicing contrasts from final to initial position to see if participants generalised in line with the established typological pattern. In typology, having a voicing contrast in final position implies also having such a contrast in initial position (Steriade, 1997). No evidence was found that supported a naturalness bias since participants were not more likely to generalise a voicing contrast from final to initial position, than from initial position to final position. However, as the author notes, this might have been
caused by the fact that the rule to exclude novel, nonconforming items in the
natural condition was structurally more complex than the one needed to exclude
novel, nonconforming items in the unnatural condition. This conflation meant
that a simplicity bias would favour the unnatural generalisation and, as Pycha
et al. (2003) highlight, it is possible that the pressure for simplicity is generally
more powerful in phonology than the pressure for naturalness.

In other experiments such confounds are not as easily identifiable and the
lack of conclusive evidence is more problematic for the hypothesis that natural
rules should be easier to learn. For example, Wilson (2003) found no learning
advantage for a nasal agreement rule over a nasal disagreement rule, despite the
former being phonetically motivated. Skoruppa and Peperkamp (2011) observed
no better learning for a version of French that exhibited vowel harmony (a natural
phonological rule) compared to a version of French with vowel disharmony when
testing French participants. In a later study by Wilson (2006) there was partial
support for the theory, since participants tended to generalise velar palatalisation
before [e] to also apply before [i], but the same was not true for generalisations
across different target consonants. In that part of the study there was no evidence
that participants generalised palatalisation from of [g] to imply that [k] undergoes
the same process, as tends to happen in typology. Similar patterns of results
are also found in studies of suprasegmental phonology. For example, Carpenter
(2010) found that, when teaching participants stress patterns, the typologically
uncommon system where stress was placed according to the rule “leftmost low
vowel, else leftmost” was learned better than the unattested “leftmost high vowel,
else leftmost” system. Yet, an earlier study by the same author did not find a
preference for the typologically common system with the rule “leftmost heavy, else
leftmost” over the unattested “leftmost light, else leftmost” system (Carpenter,
2006).

Despite this seemingly inconclusive evidence, there are some studies in re-
cent years that have produced more reliable evidence in support of there being a
substantive bias for naturalness in phonology. Crucially, these studies highlight
the importance of including generalisation and extrapolation elements in the ex-
erimental task when trying to observe naturalness effects in participants’ be-
aviour. For example, Wilson (2006) used the the name “poverty of the stimulus”
to describe the paradigm that he employed in his experiment, and subsequent

\footnote{Vowel harmony here refers to phonological assimilation processes between vowels in a phono-
logical unit (usually a word). These processes usually involve a “trigger vowel” that instigates
the assimilation process, and target vowels then adopt some number of features from the trigger
vowel.}
researchers that use the same method have stressed that using this paradigm makes it easier to draw conclusions about the rule(s) that participants have actually learned in training since they must extend these rules to novel forms and contexts (Finley & Badecker, 2007; White, 2014; White et al., 2018; Wilson, 2006). The paradigm employs a deliberate “impoverishment” of the training data, compared to the data that participants encounter during testing, meaning that participants may be exposed to only a small sub-set of the system, or to items which are ambiguous between different hypotheses. As a result, most testing phases in these types of experiments involve generalisation or extrapolation tasks. Typical generalisation tasks include withholding specific testing items, drawn from the same type/distribution as the training items, to examine if participants generalise patterns or rules to these novel items at test. Typical extrapolation tasks include the introduction of novel categories of items at test, or requiring participants to combine different categories of items and structures to create a new (often more complex) type of item during the testing phase.

In the experiments by Finley and Badecker (2007) they found evidence that learners extend a natural rule of vowel rounding harmony to both novel suffix forms and novel vowels, in line with the predictions of a substantive bias for natural phonological rules. Similarly, Finley and Badecker (2008, 2010) found that when learners are trained on data that is ambiguous between directional feature spreading of vowel harmony and the theoretically possible (Bakovic, 1999; Lombardi, 1999), but phonetically unmotivated “majority rules” system of vowel harmony (where the majority feature value spreads to segments with the minority feature value), learners infer a directional pattern. Other studies that have made heavy use of this experimental design include Martin and White (2019) and Martin and Peperkamp (2020), both of which directly compare learning of a vowel harmony versus vowel disharmony rule. Vowel harmony is typologically common and has a clear phonetic precursor in vowel-to-vowel co-articulation (Ohala, 1994) whereas vowel disharmony does not. Furthermore, the two rules are structurally equivalent in terms of the amount of primitives needed to express them, thus avoiding potential confusions with simplicity. Both studies find evidence that a vowel harmony rule is learned and generalised better than a vowel disharmony

---

4Note that Wilson (2006) found mixed results in support of a bias for phonological naturalness. In Experiment 1 participants generalised velar palatalisation from more-back vowels to less-back vowels, in line with a bias for natural palatalisation rules, but in Experiment 2 they did not systematically generalise palatalisation of voiced consonants to voiceless consonants, which would have been the prediction under a naturalness account.

5Directional feature spread of vowel harmony is motivated by processes like coarticulation and gestural overlap during speech production (Martin & Peperkamp, 2020; Ohala, 1994).
rule when employing this type of “poverty of the stimulus” design.

Leaving the domain of phonology and returning again to syntax, interesting parallels can be drawn between the concept of naturalness across these two domains. Just as phonological naturalness finds its motivation in a system external to phonology, namely phonetics, naturalness effects in syntax can be seen as being motivated by aspects of semantic structure and language processing constraints (Rosenbach, 2003; Schouwstra et al., 2017). For example, semantic structure has been put forward as the organising principle behind certain syntactic hierarchies (Culbertson & Adger, 2014; Dryer, 2018; Martin et al., 2019). Furthermore, semantic content and salience can also determine the accessibility of individual concepts, thus privileging those concepts during processing (Branigan et al., 2008; Rosenbach, 2003).

One word order pattern that has been claimed to originate in a cognitive bias for naturalness in the syntactic domain is the asymmetry of word orders in complex noun phrases. An example of a complex noun phrase would be *these three cute cats*, which consists of a demonstrative, a numeral, an adjective and a head noun. Typological data shows that a small subset of the twenty-four possible permutations of the components of a complex noun phrase make up the vast majority of orders found in the world’s languages (Cysouw, 2008, 2010; Dryer, 2018). The two most common orders are the two harmonic orders, the prenominal harmonic order Dem-Num-Adj-N is the one we see in English, and the postnominal harmonic order N-Adj-Num-Dem is found in languages like Thai. The other common orders are similar in that they also tend to order adjectives closer to the noun than numerals, and numerals closer to the noun than demonstratives (see Figure 1.2a for distribution of orders). These types of orders have been argued to transparently reflect, or be homomorphic to, the underlying hierarchical structure of the noun phrase (Culbertson & Adger, 2014; Rijkhoff, 2004).6 Several attempts have been made to derive this typological pattern using different theoretical approaches (Abels & Neeleman, 2012; Cinque, 2005; Dryer, 2018). Although these accounts differ in the mechanisms they use to derive the proposed underlying hierarchy, they all make reference to some aspect of semantics as being the organising principle behind the underlying structure. For example, Cinque (2005) and Culbertson and Adger (2014) and Rijkhoff (2004) all make reference to semantic scope as the basis for the structure in Figure 1.2b.

---

6Note that the linearisation of the underlying NP hierarchy has previously been argued to be a reflection of *simplicity* as the process of linearisation from the hierarchy to surface order requires fewer permutations for homomorphic orders (Culbertson & Kirby, 2016). However, here I take the stance that this pattern constitutes a naturalness effect, similar to iconicity, in that the underlying hierarchy is reflected in the linear order in an iconic/transparent way.
herentness has also been used to explain the ordering preferences (Culbertson & Adger, 2014; Dryer, 2018; Martin et al., 2019) by arguing that dependents that denote more inherent properties will appear linearly closer to the noun to which they ascribe those attributes. In the context of complex noun phrases like *these three red strawberries*, the adjective *red* denotes an inherent property of (at least ripe) strawberries, whereas the fact that there are three of them is not a property we ascribe to strawberries by default. The adjective *red* is therefore more inherent to the concept of strawberry than the number *three*. Thus it has been claimed that humans are biased to prefer word orders that transparently, or naturally, reflect semantically motivated hierarchical syntactic structure (Culbertson & Adger, 2014; Rijkhoff, 1990, 2004).

There are a number of recent experimental studies that support this hypothesis with regards to the order of elements in the noun phrase. Just like in the case of phonological naturalness, the studies that find a preference for these natural word order patterns make use of tasks where participants have to produce structures/items that they have never seen direct evidence for (Culbertson & Adger, 2014; Martin et al., 2019, 2020, 2024). In these experiments participants are trained on the order between nouns and individual dependents but are never given evidence for the relative ordering between different dependents. At test they have to produce complex noun phrases with two dependents, and the orders they infer between dependents tend to align with those we see in typology and are motivated by a preference for homomorphic word orders. Additionally, the same kind of word order preference is observed in silent gesture experiments where participants who are used to communicating in the vocal/auditory modality have to improvise gestural expressions for complex noun phrases. Here the order of gestures also tend to align with the homomorphic word order preference when participants express meanings like *these three striped squares* or *those two big toothbrushes* (Culbertson, Schouwstra, & Kirby, 2020; Schouwstra et al., 2017). Interestingly, these same studies find no evidence that participants’ improvised productions tend to be harmonic.

The order of subject, object and verb is another instance in which semantic content can influence syntactic structure. In the silent gesture study by Goldin-Meadow et al. (2008) participants were asked to improvise gestures for a combination of simple events such as GIRL-WAVES, WOMAN-TWISTS-KNOB and GIRL-GIVES-FLOWER-TO-MAN. The study found that, independent of the order of subject, object, and verb in participants’ native language, they tended to gesture in an order equivalent to SOV. The authors termed this the “natural order of events”, as it places more cognitively basic entities (actors and patients
Other studies find that the order of subject, object and verb can be conditioned by the semantics of the specific event that is being expressed. For example, Schouwstra and de Swart (2014) and Schouwstra et al. (2020) found that when participants improvised gesture they would gesture in SOV for extensional events (e.g. *clown rings bell*) but shift to SVO when they were gesturing intensional events (e.g. *pirate hears guitar*). This verb-type conditioning for basic word order has also been found in examinations of young signed languages (Flaherty et al., 2018) and some older sign languages (Napoli et al., 2017). Similarly, many gesture studies have investigated the difference between non-reversible (i.e. events involving an animate subject and an inanimate object) and reversible events (i.e. events involving an animate subject and an animate object), and a shift from SOV gesture order to SVO for reversible events has been found in several experiments (Futrell et al., 2015; Gibson et al., 2013; M. L. Hall et al., 2013, 2014; M. Hall-Harzi et al., 2014).

I use the term “young” here to denote these languages. This is due to current debate in the field regarding the appropriateness of previously used terms like “emerging”, since term may be interpreted as devaluing the status of these signed languages as fully-developed and functioning languages, by suggesting that they are still in the process of attaining language status.
Meir, Lifshitz, et al., 2010). Yet these studies vary greatly in the consistency with which they find this effect and, as Meir et al. (2017) highlight in their study, the pattern that they do see clearly across studies is a subject initial preference. In fact, the typological data of both spoken and signed languages show that SVO or SOV are the most common orders across both spoken and signed languages (Newmeyer, 2000). Data based on spoken languages from WALS shows that over 75% of the languages have either SVO or SOV as their default basic word order (Dryer, 2013e). These are also the majority order found in a study of constituent order across 42 signed languages (Napoli & Sutton-Spence, 2014). Meir et al. (2017) suggest that this pattern may be related to a “human first” bias where animate, and especially human, referents tend to appear early in a phrase. How much this is caused by the animacy status of the referent is a bit unclear in studies involving gesture and signed language. In such systems the human body is itself an articulator and the tendency for signers to inhabit the role of a referent with their own bodies might drive this pattern beyond what we see in spoken languages (Meir et al., 2007).

1.4.3 Category-specific and system-wide biases

What is clear from this examination of the literature is that, to properly examine the effects of biases which are grounded in either naturalness or simplicity considerations, one needs to find a way to disambiguate the two types of biases in experimental tasks. This is a general challenge when conceptualising biases based on naturalness and simplicity since they are also prone to conflation in theoretical frameworks and experiments. For example, several phonological theories are built to express natural patterns in a simple way since such patterns are so common (Moreton & Pater, 2012a). A prime example is Natural Phonology (Stampe, 1979), where acquisition of any unnatural rule requires the activation of mental inhibitors to suppress the default natural processes, arguably making grammars that include unnatural rules more complex (Tobin, 2009). A first step towards differentiating these biases is to establish at which level in a linguistic system that simplicity and naturalness are active.

Based on the formulation of the biases of simplicity and naturalness in the literature they can be reconceptualised in the following way. Simplicity can only be defined in relation to a system, or in a comparison between items. For example, we cannot determine if a system exhibits the property of word order harmony by looking at just one simple phrase. There is nothing meaningful to be said about the harmonic structure of a system if all we see is Noun-Adjective. However, if
we also see examples of Noun-Genitive or even Preposition-NP we can say something meaningful about the cross-category harmonic structure of that system. The more expansive a system is, with regards to number of categories, the more beneficial harmony becomes as an organising principle. As such, any grammar which requires more rules, or more specific rules, to capture the structure of the language can be seen as more complex. Thus simplicity relates to the systematic organisation of linguistic items within a language system. It is not a property of any individual linguistic item (Noun-Adjective), rather it is a property of the system that generates the item (XP → X...). Naturalness, on the other hand, finds its motivation in processes that are external to the domain in which its effects manifest. As such, any individual syntactic item’s naturalness status can be established by examining how its word order reflects semantic content and processing constraints. These can be judged in isolation, and do not need to be evaluated with regards to the system they belong to. For example, a complex noun phrase with the structure Noun-Adjective-Demonstrative makes for a transparent and natural reflection of underlying semantic structure. Similarly, the naturalness status of an individual phonological alternation is defined by how well it adheres to general phonetic processes (e.g. is it motivated by principles of production effort). In phonology the preference of vowel harmony is thus a category-specific bias as it is motivated by coarticulation, whereas vowel disharmony is not. With this definition, an individual linguistic item or alternation can be evaluated as natural or not without reference to the system at large.

With these differences in mind, many naturalness phenomena can be classified as category-specific. They are reflections of biases which act upon individual items or individual categories of items within language. Simplicity phenomena, on the other hand, are reflections of biases that concern the structure of a whole linguistic system, and finds their motivation from within that system, making such biases system-wide. Therefore, we might expect simplicity to have an effect in tasks that require consideration of a whole system, whereas naturalness effects are more likely to appear in tasks that require evaluation of individual items.

Having established the formulation of the two types of biases, let us return to an examination of the experimental evidence in the literature. At present, there is a significant amount of data supporting the hypothesis that people have a preference for imposing systematic organising principles that result in simpler grammars for both phonological and syntactic structure, and that these systems are easier to learn. Thus, it seems that biases which acts on a system level have clear behavioural effects during acquisition of system-wide patterns. The effects of category-specific biases are more challenging to equate across the two domains.
In syntax, there is a significant amount of experimental evidence indicating that semantic structure (such as the type of verb or animacy status of specific referents), and processing of such structure, can determine word order choice. In phonology, although the grounding of naturalness in the external system of phonetics provides a more unified foundation for naturalness in this domain, the experimental evidence is sparse. Yet syntactic literature, as well as those studies in phonology that do find support for a naturalness bias, emphasise the need to incorporate generalisation or extrapolation into the tasks in order to observe effects of both syntactic and phonological naturalness in experimental contexts.

1.5 Scale of Innovation

Based on the evidence outlined above, suggesting that system-wide and category-specific cognitive biases potentially influence linguistic behaviour in different contexts, I now explore these contexts in more depth and present them along a continuum that I call the Scale of Innovation. This scale constitutes a theoretical construct I employ to order linguistic tasks and contexts along a continuum based on the amount of innovation, or improvisation, that people need to employ during that particular linguistic task. The divisions along the Scale of Innovation are informed by three distinct sources of evidence: (i) potential pathways of language evolution and the stages at which different elements of linguistic structure may have emerged, (ii) language acquisition data, showing parallels between contexts and language acquisition, and (iii) the characteristics of experimental tasks that have been used in the past to study phenomena related to cognitive biases and language evolution. I will focus my description below on the latter two sources of evidence, and save in-depth discussion of the evolutionary perspective to the thesis Discussion in Chapter 5. Most of the experiments I draw upon in the section below are those that use some version of artificial language learning, although the scale may also be applicable to other experimental methods.

An illustration of the Scale of Innovation can be found in Figure 1.3. The scale progresses from left to right, with contexts requiring the least amount of linguistic innovation on the far left, and contexts that primarily rely on innovation on the far right. The following sections (1.5.1-1.5.5) are dedicated to detailing the properties of tasks that exist in each region of the scale, as well as providing examples of experiments in linguistics and language evolution that would fall within each of these regions.

50
1.5.1 Improvising

Starting on the far right of the scale, where the need for innovation is highest, is the region I refer to as *improvising* following similar arguments in the literature (see e.g. Schouwstra et al. 2020). Essentially, you can imagine all tasks in this region of the scale as representing some form of “blank slate” state for a communication system. In this type of contexts, there are no shared conventionalised linguistic signs to use as the basis for communication. However, even with this lack of shared signals, there are other important shared features that people may rely upon during communicative events to convey their intended meaning(s). These shared features make up the constraints and grounding for category-specific biases as they are external to the language system but still capable of influencing specific communicative events. These shared features include shared constraints on information processing and attention, as well as features related to the physiology of speech production, and the perception of speech sounds.

One important strategy which can be leveraged in improvisation contexts is to rely on transparent relationships between novel linguistic signs and intended referents. This type of transparency is usually called *iconicity* and, in spoken language, iconicity refers to some level of correspondence between the sounds used for a word and the referent it denotes. A prime example of this is the words that spoken languages use to denote the sound that various animals make. In English we have bees that *buzz*, cats that *meow* and ducks that *quack* (etc). These all bear a sound-symbolic or iconic similarity to the events they denote in the world. However, this transparency is not fully predetermined since spoken languages exhibit quite a bit of variability in the sounds they employ in these types of words. In Swedish bees *surrar* and in Japanese the sound a cat makes is ‘にゃん’ (*nyan*).
Other types of iconicity in spoken language are more subtle and instead leverage shared associations between sounds and certain concepts in the world (Blasi et al., 2016). For example, in studies looking at the famous bouba/biki effect, participants are shown images of spiky and rounded shapes and then played recordings of novel words like ‘bouba’ and ‘kiki’. They are then asked to choose which shape they think is associated with which word. A recurring pattern across these studies is that participants tend to associate the word with the low vowels (‘bouba’) with the rounded shape, and the word with the high vowels (‘kiki’) with the spiky shape (Ćwiek et al., 2022; Ramachandran & Hubbard, 2001). Similarly, there are groups of words that have phonesthemes, which are specific phonemes that when they appear together in words, denote specific abstract concepts. An example in English is the gl- onset cluster, which occurs in words like *glimmer, glow, glitter, glisten* and so on. All of which relate to a perception of a certain light/visual quality (Marchand, 1959; Nuckolls, 1999). While this phonestheme is most likely the result of an historical accident, other sound-meaning associations have been argued to reflect more fundamental iconic features of spoken language. For example, in a recent study by Winter et al. (2022), the authors investigated the association between trilled /r/ and “roughness” across 84 language families. They found strong evidence that this is a non-arbitrary association, and they argue that it is grounded in the iconic association between the acoustical discontinuity of the trilled /r/ and the discontinuity of surface texture that identify surfaces as “rough” through touch. The existence of this type of non-arbitrariness in spoken language has received more attention in recent years (Dingemanse et al., 2015; Nuckolls, 1999; Winter et al., 2023), and it calls into question how arbitrary spoken linguistic signs truly are.

In signed languages iconicity operates on visual rather than auditory similarity between signal and referent. For example, signed languages may make use of hand shape to signify the physical shape of objects, or have signs that emulate parts of actions as a sign for a specific verb (Padden et al., 2013). Signed languages also have a unique ability to make use of simultaneity and physical space to iconically express features such as grammatical relationships and syntactic category membership (Aronoff et al., 2008; Meir et al., 2007; Napoli & Sutton-Spence, 2010; Padden et al., 2013; A. Senghas et al., 2004). For example, persons can be set up in different regions in the signing space and then referred to by indicating (usually with a pointing sign) to that part of the signing space later in the utterance. Some of these types of iconicity are found more in early cohorts in instances of new language creation, indicating that languages may go from relying heavily on this type of transparency to more arbitrary representa-
tions over generations (A. Senghas et al., 2004). However, while iconicity can be used to help scaffold the association between linguistic signs and meanings, iconic sign systems are not necessarily easier to learn how to produce compared to arbitrary ones (Sato et al., 2020). This result lends further support to the idea that favoured structural features differ at different ends of the scale as iconic signals are preferred in improvisation, but this does not necessarily translate to a learning advantage for those signals.

While improvisation context are an interesting source of evidence in how individuals make use of shared features to communicate, it is extremely rare to come across, and be able to study, instances of complete language improvisation in a modern context. However, a rich source of evidence that linguists have depended on to understand how language behaves in these contexts is the documented cases of new sign languages that have been created in Deaf communities in modern times. These new sign languages include Nicaraguan Sign Language, created by groups of deaf children who were first brought together in a dedicated school for the Deaf in the Nicaraguan capital, Managua, in the late 1970s (Goldin-Meadow et al., 2015; A. Senghas, 1995; A. Senghas et al., 1997, 2004; R. J. Senghas et al., 2014). Others include smaller “village” sign languages, such as Central Taurus Sign Language in Turkey and Al-Sayyid Bedouin Sign Language in Israel, which were created within localised geographic areas where there was a high level of hereditary deafness and limited access to national signed languages (Ergin, 2017; Ergin et al., 2018; Meir, Sandler, et al., 2010; Padden et al., 2010; Sandler et al., 2005). In these contexts, researchers have been able to study the process of emerging systematicity from isolated instances of improvised gestures to conventionalised language systems, observing a development from high levels of iconicity to increasing amounts of linguistic arbitrariness over time.

A related context which could provide us with valuable insight into the structure of improvised linguistic communication is the creation of homesign systems. These systems are the inventions of deaf children who are isolated from the wider Deaf community, meaning that they have no access to an official signed language in their home. Parents of these children often favour “oralism”, which involves focusing on spoken language and lip-reading as the main forms of communication, rather than promoting the use of sign language in the home (W. C. Hall, 2017). This practice deprives deaf children of important social and linguistic input from their peers and tends to leave them isolated from their community.

Note that some iconicity that makes use of simultaneity in signed languages are in fact emergent phenomena, these seem to be the type of iconic representations that require very high proficiency in the use of multiple articulators (Slonimska et al., 2022).
As a result, these children tend to create their own manual linguistic systems, called “homesign” systems, to communicate with people in their close proximity (Goldin-Meadow, 2019). As in the cases of young sign languages, researchers have examined the development of systematicity within homesign systems as a way of understanding how linguistic structure emerges over time, even in the absence of community-wide pressures. Key insights from this research include the fact that homesign goes from holistic to compositional structure. For example, A. Senghas et al. (2010) observed how different children using homesign would produce signs that combined manner and path information in a single gesture, and later separate out these two features of the meaning to separate signs. The same pattern is also found in studies of young sign languages (A. Senghas et al., 2010) and in experimental work by A. Senghas et al. (2004), where they use silent gesture to show that there is an initial preference for holistic signs that indicate both the manner and path of an event. Furthermore, while most other language creation contexts involve a cooperative process, where a wider community needs to agree on the representation of specific lexical items/grammatical features, homesign systems are unique in that they appear to be the work of a single individual (E. M. Carrigan & Coppola, 2017). Studies examining the relative effect of parental input on the structure of homesign systems have found that the true guiding force behind the linguistic structure in these systems is the individual child, and not their frequent interlocutors’ productions (E. Carrigan & Coppola, 2012; Goldin-Meadow et al., 1984).

Approximating contexts in language evolution where there is no language model for participants to be influenced by is challenging since, unlike children who have been linguistically isolated, all adult participants come to experimental tasks with extensive knowledge of at least one language. This knowledge can interfere with an experiment that aims to study linguistic productions in a context where participants have no established linguistic system to rely on. One methodology which has proven successful in limiting the influence of participants’ native language(s) is the silent gesture, or elicited pantomime, tasks (see e.g. Goldin-Meadow et al., 2008) described previously in Section 1.4. In experiments that use this method participants who normally communicate in the verbal modality are tasked with creating novel linguistic expressions for specific meanings using only their hands. The types of patterns observed across both contexts of young signed languages, and in improvisation tasks using the silent gesture methodology, tend to be those which are favoured by category-specific biases. As previously mentioned, silent gesture studies have revealed that participants tend to improvise homomorphic noun phrase word orders (Culbertson, Schouwstra, & Kirby, 2020).
Furthermore, the conditioning of basic word order based on event type has been found in both silent gesture studies (Schouwstra & de Swart, 2014) and in some young (Flaherty et al., 2018), and some more established (Napoli et al., 2017), signed languages. These parallels in results suggest that silent gesture is an effective experimental proxy for language improvisation contexts. Furthermore, recent studies have shown that preferences which are observed in production-based tasks using silent gesture are also observed in judgement-based tasks using gestural stimuli. Studies that use this latter design show participants videos of gestures that express specific meanings, and ask participants to select the gesture video they think is best for the given meaning and find parallel results for event-type conditioning as with production-based methods (Do et al., 2022; Motamedi, Wolters, Naegeli, et al., 2022). Gestural stimuli may also be used in artificial language learning tasks, where participants are trained on a language system (Motamedi-Mousavi, 2017), but when the aim is to approximate improvisation contexts the combination of silent gesture and lack of training is the best combination.

Returning to the Scale of Innovation, these results suggest that researchers who are interested in studying category-specific biases should aim to design experiments that approximate contexts where no conventionalised language system is present. Experiments that make use of gestural stimuli in either a production-based or judgement-based task without systematic training seem to be able to act as a good proxy for this type of context as it limits the influence of native language on participants’ linguistic behaviour. Crucially, these experiments limit exposure to anything which could be conceived of as a language system during the experimental tasks, instead participants have to employ high levels of linguistic innovation to complete the tasks. This leads to activation of category-specific biases on this side of the Scale of Innovation.

1.5.2 Extrapolating

The next region on the scale is one I have labelled as extrapolating since the tasks and contexts in this part of the scale still require a high amount of innovation, but usually there is some part of the language system in place already. Contexts in this region of the scale can be thought to involve “incomplete” language systems,
in that there are some conventionalised signals but where whole categories of signals still need to be actively innovated by individuals. From an acquisitional point of view, this parallels findings that certain categories of linguistic items tend to be acquired earlier than others, meaning that the population of the linguistic system during acquisition leaves certain regions of the meaning space without an underlying system for longer periods of time (Gentner, 1982). Without clear evidence for how to treat a new category of items people may rely on similar biases in these contexts as in improvisation tasks. Yet the fact that some evidence from linguistic items of another type is available may cause participants to also rely on some more system-wide biases, relating to the systematic integration of these concepts into a grammar. As such, this region of the scale may be a place where both types of biases play a role in guiding people’s linguistic behaviour.

Looking at experimental work, extrapolation tasks include those where participants lack direct evidence of how a specific structure should be systematically expressed. In these tasks participants are required to construct an expression of that novel structure based on partial evidence and/or component meanings. Due to this lack of direct evidence these types of tasks are related to paradigms that employ the “poverty of the stimulus” method, where structures or examples of interest are purposefully withheld from participants during training (Culbertson, 2023). As discussed previously in Section 1.4.2, this method has been employed to study syntactic phenomena like noun phrase word order (Culbertson & Adger, 2014; Martin et al., 2020), phonological rule learning (Finley, 2018; Martin & White, 2019; White, 2014; White et al., 2018; Wilson, 2006) and morphological phenomena like the order of affixes (Saldana, Oseki, & Culbertson, 2021). For example, Wilson (2006) employed this methodology to test which phonological rule participants internalised when they were trained on data that could be explained by two different rules, and then given disambiguating examples during test that showed which of the two rules they had learned. Culbertson and Adger (2014), on the other hand, used a slightly different version of this methodology, they trained participants on individual combinations of nominal heads and one dependent at a time (e.g. Noun-Adjective and Noun-Demonstrative). Participants were then asked to generate complex noun phrases where both dependents were present at the same time. Nothing in the training data gave them any indication of the relative order between these dependents, and so participants had to extrapolate the order in the complex noun phrase at test. Culbertson and Adger (2014) employed the same design, but applied it to ordering of number and case marking, showing that participants’ extrapolated order matched the typological tendency for number morphemes to be closer to the stem than case morphemes in
languages where these morphemes are stacked. In these studies, the results indicated that participants were guided by category-specific biases favouring natural phonological rules, and those orders which transparently reflected the underlying semantic structure for both word order and morpheme order.

While the methods in these experiments vary somewhat from one another they all rely on extrapolation at test by asking some version of the question “what do participants do when they have no direct evidence of how to perform the task they are being asked to perform?” In the case of the phonological research, this is done by providing disambiguating examples at test, in the syntactic and morphological examples it is done by asking participants to generate a more complex structure than what they saw during training. But some extrapolation can also be conceptualised in a slightly different way, which places the task right at the edge of the extrapolation and generalisation regions of the Scale of Innovation. In these types of contexts, participants are tasked with creating a previously unseen structure, in a context where there is no direct evidence for how that structure should be expressed. However, it is crucially different from the examples from the syntactic and morphological examples outlined above in that there is abstract evidence for the structure that participants need to produce, but this evidence comes from structures/phrases of a different type than the target type. Examples of studies that employ a similar kind of extrapolation task are Wang et al. (2023) and Cook (1988), where the authors examine how participants extrapolate word orders from one phrase type to another in an artificial language learning task. In these studies, participants are trained on the order between one type of phrasal head and one dependent and then, at test, they encounter a new type of phrase and must decide to either harmonise the head and dependent of this new phrase with the one seen in training, or to use a different word order for the new phrase type. This type of extrapolation can be defined as examining cross-category extension of patterns or rules in language.

The extrapolating region of the Scale of Innovation includes a group of tasks that involves the production of linguistic structures in contexts with limited evidence for the organisation of that linguistic structure. This type of task therefore involves both learning of a language system, and innovation within that language system, making it a possible site of competition between category-specific and system-wide biases that are misaligned and combined influence of these bias types when they are aligned. Researchers who are interested in comparing the strength of competing cognitive biases may therefore consider using extrapolation tasks to study these phenomena. Depending on the phenomena they are studying, they may choose to use tasks of the type where the building blocks
of a target structure are provided but never shown in combination (Culbertson & Adger, 2014; Martin et al., 2020; Saldana, Oseki, & Culbertson, 2021) or use tasks where the only available evidence for the target structure comes from a structure of a different type (Cook, 1988; Wang et al., 2023).

1.5.3 Generalising

Next to the extrapolating region of the scale, where we find contexts that require people to innovate novel structures from partial/indirect evidence, is a region which includes contexts that require generalisation. In this type of context individual categories show expansion of meanings within each class, such that for example, more and more nouns are created to sub-divide a meaning space that was previously covered by a single noun, generating more precise meanings. The extent to which different languages do this across different categories depend on the specific communicative needs of that culture. For example, there is a rich literature examining the evolution of colour name systems across the world’s languages. A recurring theme in evolutionary studies of this topic is that languages tend to evolve/acquire names for colours in a specific sequence, black and white, then red, green and yellow (either order), blue, brown, purple, pink, orange, and grey (Regier et al., 2007). This gradual lexical enrichment of the semantic space of colour may serve as an exemplar of how other semantic spaces show increasing subdivision of the meaning space.

As lexical categories expand language users are faced with the choice of applying the same rules to these new category members, for example if the same plural formation is used for all new nouns, or the same conjugation is used for all new verbs. The choice here is one between conforming to the current grammatical patterns of the system, or to innovate some novel way of treating these linguistic items. Looking at acquisitional data for patterns of generalisation there is a wealth of research examining how children generalise (i.e. extend within a category) morphological rules to novel lexical items. One example is the classic *wug* test (Berko, 1958), which focused specifically on children’s tendency to apply English plural morphology to new nouns, showing that children show an early awareness of the rules of plural morphology. This same type of task has also been conducted with other word types, for example looking at verb conjugation (Berko, 1958; Prasada & Pinker, 1993). What is interesting about people’s generalisation abilities with reference to the Scale of Innovation and language typology is if generalisation tendencies which are in line with cognitive biases are more robust, and if the type of bias (system-wide vs category specific) matters
for this robustness. Generalisation tasks can, on the one hand be seen as a “complete the system” task by participants, in which case we would expect high levels of system-wide biases to influence behaviour, but they could also be considered as another context (like extrapolation) where part of the system is missing and people need to innovate their linguistic expressions.

Experimental studies that make use of generalisation can be found in both the syntactic and phonological literature, and it is common to see a combination of novel and known items during testing in many artificial language learning experiments. Two studies that use a combination of familiar and novel items in the testing phase (thus incorporating generalisation) are Pycha et al. (2003) and Martin and Peperkamp (2020). Both of these studies compared the learnability of harmonic and disharmonic phonological rules in an artificial language but found diverging results. In Pycha et al. (2003) there was no evidence that the category-specific preference for phonological harmony impacted participants’ learning, whereas Martin et al. (2020) did find better learning for their harmony rule. Neither study reports a direct comparison between the novel vs familiar test items, but Martin and Peperkamp (2020) notes that performance was more stable for novel test items, suggesting that having to extend the pattern to novel items somehow cemented the rule learning from training.

In addition to these phonological learning studies, there is also experimental and computational work which simulates the repeated nature of the language learning context where generalisation is required. This research examines situations involving *iterated learning*. In these experimental tasks participants are trained on a language system (either with specific starting properties or randomly generated from a selection of syllables) and then have to reproduce those languages at test. Whatever they produce is then used as the training data for the next participant in the experiment. Usually participants are arranged in diffusion chains and the experiment runs for a specific number of generations (Smith et al., 2003). The same type of set-up has been used for artificial agents as well, the benefit of such experiments is that you can more precisely modify features such as the information bottleneck (i.e. how much of the data from the previous generation that the new generation is exposed to) and many more chains/generations can be run than in an experiment with human participants. Results from these experiments show how biases for efficiency/compression of linguistic systems clearly emerge over generations, sometimes causing the artificial languages to become degenerate if there is no simultaneous pressure for expressivity/communication (Kirby et al., 2008, 2015).\(^\text{10}\) What classifies these studies as residing in the gener-

\(^{10}\)Degenerate languages here refer to languages where every meaning is expressed by the
alisation region of the Scale of Innovation is the use of the transmission/learning bottleneck. This bottleneck constrains the amount of data that each new generation of users are given during training and helps simulate natural language learning contexts, as no real language learner is ever exposed to the full structure of their target language. The result of this constraint is that participants and agents slowly change the language over generations to be more robust to the challenge of generalisation, such as making the language more regular and more compositional. The key take-away from these studies is that it shows how weak individual biases (instantiated as Bayesian priors in the computational models) and the need to generalise can give rise to strong structural skews in favour of specific structural features (Kirby et al., 2007; Reali & Griffiths, 2009; Thompson et al., 2016). This is caused by the continuous application of the biases which are present during language learning and generalisation.

These examples show that, as with the extrapolating region of the scale, the generalising region also involves tasks requiring production of linguistic items in contexts with limited direct evidence for how that structure should be expressed. The level of innovation involved in generalisation tasks is, however, quite limited, and given that Wang et al. (2023) found that the tendency to extrapolate harmonic structure across phrase types as being dependent on the similarity between the linguistic items in the phrases (more similar items meant more harmonic transfer from one phrase type to the other), we might expect that system-wide biases are generally even stronger in generalisation tasks since they involve pattern extension between members of the same linguistic category (assuming that items that are members of the same linguistic category tend to be more similar than members of different linguistic categories). The preference we see for regularity and compositionality in the iterated learning studies outlined above may be a reflection of this. It is however possible that some category-specific biases are strong enough to influence behaviour in such contexts. Experimental tasks which make use of generalisation are thus most likely to reveal system-wide biases, but may also be affected by certain category-specific biases.

1.5.4 Regularising

Following these three regions on the Scale of Innovation: improvising, extrapolating, and generalising, comes contexts where newly established signals and categories become conventionalised and regularised across the wider population.
of language users. In this type of context, the level of innovation involved during language use is generally lower than in the previous contexts, which concerned the expansion of the the meaning and expression space of language. Instead, the innovation that is present in contexts where meanings are established is more related to the loss of random or unconditioned variability in the wider language system. I call this region the *regularising* region. This type of context can, to some extent, be observed continuously in all languages of the world through processes of language change as new words are borrowed, coined and fall out of use over generations. Regularisation is thus, to some extent, a community process as it involves navigating the agreement for how to express meanings in a systematic way across a population of language users, a process sometimes referred to as “self-organisation” (Hurford, 2012).

Interestingly, community structure and population size seem to matter when it comes to patterns of language variability. For example, larger communities seem to regularise and eliminate unpredictable variation from their language more readily than smaller communities. This has been observed in artificial language learning tasks that manipulate the number of communicative partners that participants have (Raviv et al., 2019). Languages with a smaller population of users seem to develop and maintain various forms of variability/idiomsyncreasy compared to languages with a larger community of users (Meir et al., 2012). This can also express itself in the complexity of systematic patterns within the language, such that languages with larger population sizes tend to have simpler, more uniform, morphological systems, with fewer case distinctions and less complex conjugations (Lupyan & Dale, 2010). Similarly, when comparing the structure of a village sign language like Al-Sayyid Bedouin Sign Language and a more widespread Deaf community sign language like Israeli Sign Language, studies find that there is more inter-signer variability in terms of both the identity of lexical signs and more sub-lexical variation (e.g. variation in sign placement) in Al-Sayyid Bedouin Sign Language (Meir et al., 2012). This suggests that increasing regularity in language at a number of levels is a process connected with large and more loosely-connected social structures, a situation which requires that there are enough people using language, and that there exists competing variants within this population of language users.

The precise mechanism underlying people’s tendency to regularise is currently

---

11 The findings that the social structure of a society impacts morphological complexity are not always consistent. Recent work using a large-scale investigation of grammatical structures across languages in Grambank data found little evidence in support of this theory (Shcherbakova et al., 2023).
a topic of active debate. Some argue that regularisation is the result of a failure by participants to properly encode the distribution of variation that they are exposed to during training (Hudson Kam & Newport, 2009). Others claim that there is evidence that participants encode the variability with high levels of accuracy, but generate a more consistent system when they themselves produce the system at test (Ferdinand et al., 2019; Perfors, 2016; Perfors, 2012; Saldana, Smith, et al., 2021). For example, Ferdinand et al. (2019) found that participants were able to encode distributions of variability with high levels of fidelity, but still showed regularisation in their own productions of the systems they were trained on. This pattern applied for both linguistic and non-linguistic stimuli (words and coloured marbles), although it appears stronger with linguistic stimuli (Ferdinand et al., 2019). Furthermore, it seems as though different types of variability, for example predictable versus unpredictable variation, may be regularised by different mechanisms. The former type of variability may be reduced due to production constraints while the latter is affected more by learning constraints (Keogh et al., 2022).

Experimental work using artificial language learning capitalises on this regularisation tendency by examining whether some linguistic patterns/features are more likely to be regularised by participants than others. This method is usually simply referred to as regularisation, although it has also been called the Mixture-Shift Paradigm (Culbertson et al., 2012), and is situated close to the left-hand side of the Scale of Innovation. These types of tasks rely heavily on learning an existing language system, but the target systems that participants are trained on contain some element of variability (Culbertson et al., 2012). For example, there may be two or more competing determiners or plural markers in the language (Hudson Kam & Newport, 2005, 2009), or there may be several basic word orders used across the language (Motamedi, Wolters, Schouwstra, & Kirby, 2022). Usually, one of these variants is the majority variant and appears in more training trials than the minority variant(s). Experiments that use this methodology to examine cognitive biases rely on the idea that favoured variants (i.e. those that satisfy our biases) will be more likely to be used by participants in their own productions/selections during the test phase, compared to disfavoured variants (Culbertson, 2012; Culbertson et al., 2012). The extent to which participants extend the use of some variants over others is taken as evidence of their preferences (Culbertson, 2012).

During these types of regularisation tasks, participants try to build up a solid model of the language system they are learning and so system-wide biases which influence the organisation of that system are often found in these experiments.
These biases include preference for regularity, which may cause participants to eliminate competing variants from the language, or make them introduce a conditioning context for the variability which transforms previous unpredictable variation into predictable variation, and biases for other types of systematicity, like harmony. Due to this, regularisation tasks are especially suited to studying biases that affect system-wide organisation of language, but may also be combined with elements of other contexts such as novel testing items/testing categories. While generalisation and extrapolation experiments have been used in studies of cognitive biases in phonology, there have been no experiments that make use of the type of variable training that we see in regularisation experiments looking at syntactic phenomena. Employing this type of design in future studies of phonological patterns could help provide a more nuanced examination of learning preferences for phonological phenomena.

1.5.5 Memorising

The final region along the Scale of Innovation represents an idealised type of language learning contexts, where people are able to fully memorise a complete language system. This region on the scale represents an extreme version of the stage that all languages of the world reside in as fully-formed communicative systems (i.e. the system is available to the learner), but is crucially different from natural language acquisition contexts as no learner is exposed to a full language system, nor are they capable of memorising every language item they experience. However, investigating individuals’ behaviour in linguistic tasks that reside in this region of the scale may still provide fruitful insights into biases that are active in natural language learning by investigating if some language features/systems are easier to memorise than others. Such differences in learning behaviour could translate to real language acquisition contexts by making learners more likely to entertain memorable grammars as their target system.

As the section about the regularising stage showed, some system-wide biases like variability reduction and harmony, are found clearly in learning tasks that include some kind of variability during training. But some of these biases can also be observed in the kind of “pure” learning contexts that fall into the memorisation region of the scale. One type of experimental task which sits firmly at the left-hand side of the scale are so called “ease of learning” experiments. These experiments tend to be artificial language learning experiments where researchers are interested to see if languages with some specific property or structure are learned better/faster than languages with some different property or structure in
contexts that require no innovation. For example, some studies use this methodology to compare the learnability of different basic word orders (Tabullo et al., 2012; Tily et al., 2011). A related methodology, adaptive tracking (Schumacher et al., 2014), uses a different type of training vs testing set-up, so that instead of there being a distinct training and testing phase participants are immediately given testing trials but continuously get feedback on their answers. This type of methodology means that there is no room for innovation in the testing phase, simply the speed with which participants acquire the language through trial and error is measured. This method has been used to examine, for example, the learnability of different morphological systems (Wagner, 2021). With these two different learnability methods the speed with which participants acquire different languages, or the accuracy with which they produce or judge the grammaticality of phrases from their target language, is used to measure their learning. This type of experimental design is thus useful for testing hypotheses that some language features provide a learning advantage over other features. For example, some studies examining the system-wide preference for harmony using ease of learning designs found evidence that harmonic languages are easier to learn in this type of context, compared to non-harmonic ones (Christiansen, 2000). However, using this type of experimental design can be problematic as it may lead to ceiling effects if the learning task is too easy. Given the fact that humans have the ability to learn any human language it is perhaps unsurprising that the miniature artificial languages that people are trained on in experiments may prove too easy, especially when these systems are presented in full to the participants. When a ceiling effect occurs it potentially obscures differences in learnability, making it hard to evaluate if the feature of interest provided a learning advantage or not (Tal & Arnon, 2022).

As in the regularisation tasks described earlier, the types of biases which are generally observed in these memorisation/“pure” learning tasks relates to the learnability of a full language system, and therefore tend to be system-wide biases. These include effects of harmony and variability reduction, both of which affect the systematic structure within a language and can aid participants’ ability to quickly establish a full model of the language system. Researchers who are interested in exploring the effect of such biases should therefore consider using experiments that rely heavily on learning to study these phenomena, but also be aware of the risk of ceiling effects if the task is too easy for participants.
The Scale of Innovation is useful both as a way of formalising our understanding of how experimental tasks and contexts invoke different mechanisms, and as a tool for experimental design when examining cognitive biases. Cognitive biases that fall into the category-specific group are predicted to mainly be active in high-innovation contexts, and so researchers who are interested in examining the behavioural effects of such biases should consider designing or using existing experimental methods that sit at the right-hand side of the scale. Furthermore, they should consider how often high-innovation contexts appear in natural language settings in order to estimate the potential impact that the biases can have on language structure. Biases that fall into the system-wide group are predicted to mainly be active in low-innovation contexts, and so research focusing on such biases should consider methods that are situated on the left-hand side of the scale. These contexts include learning and regularisation tasks, when much of the language system is in place and participants have to acquire and reproduce that structure.

The strength of the different bias types may influence how much each “bleed” into the other’s space on the scale. Some category-specific biases may, for example, also influence behaviour further down the scale in more low-innovation contexts if they are particularly strong. Similarly, while system-wide biases are predicted to be especially strong in learning contexts, generalisation and extrapolation contexts also tend to involve some level of system-learning. Therefore, such contexts might see a combination of category-specific and system-wide biases. Establishing where the cut-off point for bias influence is along the scale, if this differs between sets of biases, and between linguistic domains and modalities are interesting research areas to pursue with this scale in mind. It is exactly these kinds of questions that sparked the research you will find in this thesis. On the back of the proposal of the division among cognitive biases, and the way these align with the Scale of Innovation, I now turn to a case study which examines how these different bias types influence typological structure. I focus on where along the Scale of Innovation we see effects and competition between category-specific and system-wide biases that target the same linguistic phenomenon, namely typological patterns in noun phrase word order.
1.6 Case study: competition between system-wide and category-specific biases in nominal word order

Section 1.4 established the motivation for the division of cognitive biases into category-specific and system-wide biases. Based on the review of the literature, I then proposed the Scale of Innovation (Section 1.5) as a way to categorise tasks and contexts based on how much innovation they require to complete, and suggested that the reason we see differences in results across experiments that aim to examine category-specific and system-wide biases is because these biases tend to mainly be active at different ends of this scale. This means that we need to consider the type of bias we are examining when designing experimental tasks to give us the best chance of observing the effects of that bias. A further difficulty arises when we want to compare the effects of category-specific and system-wide biases, and that is that these biases often align with each other. The alignment between system-wide and category-specific cognitive biases can be seen in work by Carol Padden and colleagues (Padden et al., 2013, 2015) where they identify a feature in signed languages which they call “patterned iconicity”. This is a systematic strategy in hand-shape variation used in many signed languages to distinguish specific semantic categories of items. For example, some signed languages may use “handling” hand-shapes to denote a class of man-made instruments/tools, whereas other signed languages instead systematically use “instrument” hand-shapes for this same category of items. Handling hand-shapes involves shaping the hands as if the person is using a specific tool, e.g. holding the handle of a hammer and performing the action of hammering to signify the meaning of hammer. Instrument hand-shapes instead uses the hands to make up some notable shape of the item itself, e.g. representing the meaning of comb by using the fingers of the hand to represent the teeth of a comb as it is moved in a combing motion close to the hair/head. This strategy of consistently using a specific type of hand-shape based on the type of noun that is being denoted has the hallmarks of a category-specific preferences in how the distinction between this category of items is grounded in the availability of the human body as an iconic representation of both items and human action. By contrast the distinction between types of hand-shapes used across different languages to signify these meanings represents the system-wide effects of systematic organisation of items in a wider language system. The use of this type of patterned iconicity in signed languages can thus be seen as a result of the way the visual modality allows users to satisfy
both category-specific biases to represent items iconically, and system-wide biases
to use diverging iconic strategies to signify meaningful semantic class distinctions
in the language system.

Similarly, if we return to the data on word order in complex noun phrases, the
two most common word orders across all the world’s languages are those which
are both harmonic (satisfying the system-wide bias) and homomorphic (satis-
fying the category-specific bias) (Dryer, 2018). This makes it hard to evaluate
the individual contributions that category-specific and system-wide biases each
make to this kind of typological pattern. Is the prevalence of Dem-Num-Adj-N
and N-Adj-Num-Dem orders caused by equal contribution from both biases, or
is one stronger than the other? Identifying instances where the two biases com-
pete is thus key to disentangling the contributions of these bias types. Finding
such instances and studying them could help answer open questions regarding
the relative strength of the biases and in which contexts the effects of both bias
types can be observed. As Pycha et al. (2003) highlighted, and as the results
in Schouwstra et al. (2020) suggest, it is possible that some system-wide biases
are much stronger structure shaping biases than category-specific biases. There
are several possible reasons for why this might be. For example, system-wide
biases may have influenced language structure more than category-specific biases
due to the fact that they are active whenever learning takes place (Garrod et al.,
2010; Kirby et al., 2008; Motamedi et al., 2019; Schouwstra et al., 2020). This
would potentially give system-wide biases more opportunities to influence lan-
guage structure. Studies that find effects of category-specific biases, on the other
hand, emphasise the role of extrapolation or generalisation in the experimental
tasks. These are both tasks that require some level of linguistic innovation, as
part of the language system is withheld from participants during training. Lin-
guistic innovation will be higher in tasks where a learner lacks direct evidence for
the structure which they are producing. Once a syntactic rule for a particular
word order, or application of a certain phonological rule to a specific structure has
been established, such evidence might very efficiently suppress any influence from
category-specific biases. Similarly, contexts of “pure” innovation (improvisation)
are very rare in the present-day language landscape. Whereas system-wide biases
would have a continuous effect at each stage of communication and learning.

To test the individual contribution of category-specific and system-wide biases
on typology, and under what conditions each type of bias is active, my research in
this thesis focuses on a case study related to word order patterns in noun phrases.
In the following sections (1.6.1-1.6.3) I identify the typological patterns of interest
to the case study, namely the order of adjectives and genitives in relation to nouns.
I highlight experimental evidence from previous research that supports the notion that there are cognitive biases in play for these dependent types, and I outline the underlying motivation for the existence of the category-specific biases that affect adjectives and genitives.

1.6.1 Order of noun and adjective

The first dependent type I will examine is adjectives. In English noun phrases the default position for attribute adjectives is prenominal, as in the phrase *bright light*. The adjective *bright* ascribes an attribute to the head noun *light*. Adjectives in English may also appear postnominally in certain specific situations, such as when they act as predicative adjectives (e.g. in the phrase *Nik is funny*), or when modifying indefinite pronouns (e.g. *I heard something extraordinary*) (Cinque, 2010). Generally, adjectives display a typological tendency for appearing postnominally (Dryer, 1992, 2013a, 2018; Greenberg, 1963, see Table 1.2 for data on noun and adjective order). As a result, adjective-noun ordering does not correlate with the order of verb and object, nor with many other head-dependent pairs that are otherwise correlated across typology (Dryer, 1992).

It has been argued that this is caused by there being an individual ordering preference (a category-specific bias) affecting the order of adjectives, which makes them more likely to appear postnominally (Culbertson et al., 2012; Dryer, 2018). For example, participants in Culbertson et al. (2012) were least likely to regularise an artificial language which was both non-harmonic, and where the adjective preceded the noun, compared to all the other language systems participants were trained on. The authors took this to indicate that, in addition to a harmony preference, there was also a preference for postnominal adjectives. Drawing on evidence from Kamp and Partee (1995), Culbertson (2012) argues that this postnominal preference originates in the fact that many adjectives can only be properly interpreted once the context of the noun has been established. While a word like *big* can be interpreted in an absolute sense, i.e. denoting a big object compared to some representational standard, in phrases like *big butterfly* our concept of the average size of a butterfly relativises the meaning of the adjective and, despite *big* denoting a much smaller size than it does in a phrase like *big elephant*, we can still interpret the meaning accurately in both of these phrases once the noun is presented. Dryer (2018) adds to this argument by highlighting that the majority of the most common adjectives across languages (for example, scalar adjectives like *good, big, small, etc.*) are of this type, where the meaning of the adjective is relativised by the noun. These common adjectives may then
act as “attractors” which generalises the postnominal preference to all adjectives. There are some experimental results that support the notion that prenominal constructions including this type of adjective can delay processing and potential referent identification since the adjective cannot be interpreted until the relativising material (i.e. the noun) is expressed (Wienholz & Lieberman, 2019). This is also supported by research showing that speakers of languages with canonically postnominal adjectives tend to produce fewer redundant adjectives in a target search task, suggesting that postnominal adjectives potentially have an efficiency advantage (Rubio-Fernández, 2016).

Table 1.2: Typological counts of the order of nouns with adjectives in spoken and signed languages (Coons, 2022; Dryer, 2013a).

<table>
<thead>
<tr>
<th>Adjective</th>
<th>N (Spo.)</th>
<th>N (Sig.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postnominal</td>
<td>879</td>
<td>16</td>
</tr>
<tr>
<td>Prenominal</td>
<td>373</td>
<td>10</td>
</tr>
<tr>
<td>Other</td>
<td>110</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 1.3: Typological counts of the order of nouns with genitives in spoken and signed languages (Coons, 2022; Dryer, 2013c).

<table>
<thead>
<tr>
<th>Genitive</th>
<th>N (Spo.)</th>
<th>N (Sig.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postnominal</td>
<td>468</td>
<td>2</td>
</tr>
<tr>
<td>Prenominal</td>
<td>685</td>
<td>22</td>
</tr>
<tr>
<td>Other</td>
<td>96</td>
<td>10</td>
</tr>
</tbody>
</table>

1.6.2 Order of noun and genitive

The second noun phrase dependent I will focus on in my thesis is genitives. In possessive noun phrases the genitive element is the possessor of the item denoted by the head noun. For example, in the phrase *Henry’s house* the genitive element is the proper noun *Henry*, and *house* is the head noun, making this a prenominal genitive construction. Similarly, in of-possessives in English like *roof of the building* the genitive element is the possessor *the building* and *roof* is the head noun, making this a postnominal possessive construction. These genitives are not to be confused with other possessive constructions, such as predicative

---

12The scalar adjectives used as examples above are, in fact, one of the set of adjectives that regularly precede nouns in languages like French, Italian, Spanish and Portuguese (i.e. Romance languages). Generally, these languages are classified as having postnominal adjectives as their default adjective position in instances of attributive adjectives (Dryer, 2013a). Many analyses of adjective order in Romance language analyse prenominal adjective placement as the result of phrasal movement, with the adjectives still being generated in postnominal position (see Alexiadou, 2001; Cinque, 2010). The two possible placements for these adjectives in Romance languages also tend to imply a difference in intersectional meaning. For example, the Spanish adjective *bueno* (‘good’) tends to have an intersective reading when postnominal, meaning that a phrase like *un cocinero bueno* (‘a good cook’) means someone who is both a cook and a good person. In contrast, *un buen cocinero* is non-intersective, signifying the meaning that the person cooks well. As such, placement variation in these languages signal a distinct meaning alternation (Alexiadou, 2001).
possession. These possessive constructions include phrases like *Aislinn has a mandolin* and *Elizabeth owns a sewing machine*. Here possession is indicated by distinct verbs signalling possession, *have* and *own* respectively. Such sentences do not contain genitives, and tend to conform to the overall basic word order of the given language (Subject-Verb-Object for English).

When examining the typological spread of the order between genitives and nouns in spoken languages the skew towards prenominal genitives is slightly weaker compared to the skew we saw for postnominal adjectives (see Table 1.3). The data from signed languages show a stronger skew towards prenominal genitives, but there is also a high number of languages which have variable word order for genitives in the signed language sample (Coons, 2022). In signed languages, just like in spoken languages, possession can be indicated in a number of ways. One common strategy when expressing attributive possession is through juxtaposition of the possessor and the possessed. In these structures the sign for the possessor and the possessed simply follow each other in sequence, without any additional possessive markers or intervening elements (Zeshan & Palfreyman, 2017). This strategy can be found in a number of signed languages, such as Adamorobe Sign Language (a village sign language in Ghana), Kata Kolok (a village sign language in Bali), Jordanian Sign Language (LIU) and American Sign Language (ASL) (Hendriks et al., 2008; Perniss & Zeshan, 2008). Examples 3-5 show examples of possessive juxtaposition in both Jordanian Sign Language and American Sign Language. The use of genitive case marking, parallel to what we see in many spoken languages is, so far, unattested in signed languages (Zeshan & Palfreyman, 2017). Instead, languages like Catalan Sign Language (LSC) make use of dedicated possessive pronouns in NP constructions, often paired with possessive “linker” signs, to indicate possession (Quer et al., 2008) (see example 7). In addition to these expressions of attributive possession, signed languages also have a number of ways to express predicative possession, including dedicated verbal signs to indicate possession and using spatial morphology through indexing (see 8 and 9 for examples from Flemish Sign Language (VGT) that illustrate both strategies) (Zeshan & Palfreyman, 2017).

(3) LANGUAGE DEAF
    language deaf
    ‘The language of the Deaf’

    (LIU: adapted from Hendriks et al. (2008) example 9a)
(4) MOHAMMED PROBLEM NOT-MY-BUSINESS
Mohamed problem not-my-business
‘Mohammed’s problem is none of my business’

(LIU: adapted from Hendriks et al. (2008) example 9b)

(5) SISTER NOSE
sister nose
‘Sister’s nose’

(ASL: adapted from Chen Pichler and Hochgesang (2008) example 17)

(6) CAR HEADLIGHTS
car headlights
‘Headlights of the car’ / ‘Car headlights’

(ASL: adapted from Chen Pichler and Hochgesang (2008) example 19)

(7) BOOK DE TEACHER
book linker teacher
‘The teacher’s book’

(LSC: adapted from Quer et al. (2008) example 2)

(8) HAVE CAT INDEX-1
have cat index-1SG
‘I have a cat’

(VGT: adapted from de Weerdt and Vermeerbergen (2008) example 27)

(9) INDEX-2 BOOK INDEX-2
index-2SG book index-2SG
‘You have a book’

(VGT: adapted from de Weerdt and Vermeerbergen (2008) example 31)

A possible explanation for why genitives are generally preferred in prenominal position can be found when we consider the type of entities that usually take on the role of possessors. Although part-whole relations may involve inanimate possessors, prototypical possessors tend to be human, or at the very least animate (Piotrowska, 2021), and as was briefly mentioned in relation to studies examining basic word order preferences (see Section 1.4.2) animate entities tend to be linearised early in phrases. This general tendency for animate entities to hold some privileged position in syntactic structure has been observed across typology.
Dahl & Fraurud, 1996; Yamamoto, 1999). Rosenbach (2003) has specifically attributed such cases to a type of category-specific bias that reflects semantically motivated conceptual accessibility and syntactic processing constraints in the linear order of lexical items. Due to the incremental nature of syntactic processing, production is fundamentally affected by the accessibility of items: highly accessible items are retrieved faster and, consequently, are also processed and available for linearisation before items that are less accessible. This means that highly accessible items might be privileged for certain positions or functions in the syntactic structure (Branigan et al., 2008; Kempen & Harbusch, 2004; McDonald et al., 1993). Evidence for how concept accessibility influences syntactic structure includes, for example, priming studies (Bock & Loebell, 1990; Branigan & Feleki, 1999). These studies also show that item-internal semantics, rather than just the item’s discourse status, can influence accessibility. For example, Bock and Warren (1985) argue that conceptual accessibility is boosted by the level of animacy that the item encodes since animate entities have a higher number of pathways associated with them (i.e. they can enter into more relations with other objects) making them easier to retrieve during processing.\(^{13}\) The animacy status of an item can thus influence syntactic structure in two ways, either the high accessibility of animate entities predisposes learners to assign them specific functional roles, such as subjects, which tend to be realised initially. Alternatively, animacy status directly affects positional processing so that animate entities are preferred early in the phrase (Branigan et al., 2008).

Several experimental studies examining the order of subject, object and verb in a multitude of languages have found support for the first (Bock & Warren, 1985; Bock et al., 1992; McDonald et al., 1993) or the second theory (Branigan & Feleki, 1999; de Smedt, 1994; Prat-Sala & Branigan, 2000), and some even support a hybrid effect where both levels of processing are simultaneously influenced by animacy (Branigan et al., 2008). Further evidence that lends support to the second of these theories relate to the order of possessor and possessed in genitive noun phrases. Studies of several languages (including English, Dutch, Swedish and modern Low Saxon) with variable order of possessor and possessed have found that animacy is a strong predictor of the specific possessive construction that is chosen in any given phrase, even when other factors such as weight and topicality are controlled for (Dahl & Fraurud, 1996; O’Connor et al., 2013; Piotrowska, 2021; Rosenbach, 2005, 2008; Strunk, 2004; van Bergen, 2011).

\(^{13}\)Prat-Sala and Branigan (2000) label the intrinsic accessibility of an item due to its semantic properties the “inherent accessibility” and this can, in turn, be boosted by “derived accessibility” which refers to the item’s accessibility in the relevant context.
The effect of animacy could potentially contribute to both the typological tendencies I have outlined. In addition to the fact that certain adjectives are processed more efficiently when they are postnominal, the noun is the only element which can represent an animate entity in this construction. Therefore the grounding of the preference for prenominal genitives, where the genitive tends to be an animate entity, may also contribute to the preference for postnominal adjectives since this order would also place animate entities early in the phrase.

There are many other pieces of evidence that support the idea that animate entities generally occupy a special place in language. For example, Dahl (2008) argues that the conceptual dominance and high accessibility of animate entities enables grammatical systems to draw distinctions for case marking according to divides of animacy. He claims that this is what gives rise to systems of differential object marking, inanimate agency avoidance in transitive sentences and the rare, but attested, differential agent marking systems (for example, see Aissen, 2003; Fauconnier, 2011; Leonetti, 2004). Furthermore, Dahl (2008) points to the default of animacy in our conceptual organisation as the reason for why, historically, animate marking tends to spread to inanimate/neuter referents (Croft, 1994). This idea is echoed in the argument by Rosenbach (2003), that natural syntactic structures are more likely to be generalised in grammatical systems, suggesting that we should see the natural “default” of animacy spreading to other aspects of the syntactic system. Further supporting evidence for the general importance of animacy to our cognitive system comes from developmental literature where many studies show how human vs non-human and animate vs inanimate are among the first categorical distinctions that humans learn to recognise (see Opfer & Gelman, 2011). Animate entities also receive priority during visual processing (New et al., 2007), are recalled better (Nairne et al., 2013) and simple changes to aspects of inanimate stimuli, such as movement or texture, can induce perception of animacy in humans and non-human primates (Giorgio et al., 2017; Scholl & Tremoulet, 2000; Tsutsumi et al., 2012).

1.6.3 Biases in competition

Having outlined the individual data for these two dependents, showing a postnominal adjective tendency and a prenominal genitive tendency in the typology of spoken and signed languages, an interesting picture emerges when examining genitive and adjective ordering relative to nouns in conjunction with other noun phrase dependents (Dryer, 2013c). For example, noun phrase dependents like numerals seem to show ordering preferences of a similar type to the ones we ob-
serve for adjectives and genitives. Numerals have an overall tendency to appear postnominally in spoken languages, but this preference is not as strong as for adjectives and does not generalise to signed languages (see Table 1.4). Furthermore, when we look at adjective, numerals, and noun ordering in combination we see that the order which places both numerals and adjectives after the noun is vastly more common than any of the alternative orders, and the second most common order is the prenominal harmonic order. (see Table 1.5). This shows that other noun phrase dependents pattern according to the bias for harmony. Yet when examining the correlation between Adj-N and Gen-N placement in a similar way (see Table 1.6) the postnominal harmonic order is as common as the disharmonic Gen-N-Adj order (Dryer, 2013a, 2013c). This pattern for adjectives and genitives might reflect an instance in which the natural bias for mirroring processing constraints in syntactic structure is biasing learners to keep these two dependents split across the head noun, a preference which is not present for numerals. Whereas, at the same time, the simplicity bias in favour of harmony would predispose learners to harmonise such a pattern over time. This competition between two bias types make adjectives and genitives the perfect test-case for an examination of the contribution that different word order preferences make to typological patterns. The competition allows for a comparison of the strength of biases, and an examination of the types of contexts in which we see each type, or both types, of biases influencing people’s linguistic behaviour.

<table>
<thead>
<tr>
<th></th>
<th>N (Spo.)</th>
<th>N (Sig.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postnominal</td>
<td>608</td>
<td>11</td>
</tr>
<tr>
<td>Prenominal</td>
<td>479</td>
<td>15</td>
</tr>
<tr>
<td>Other</td>
<td>67</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 1.4: Typological counts of the order of nouns with numerals in spoken (spo.) and signed (sig.) languages (Coons, 2022; Dryer, 2013d).

1.7 Thesis outline

The main part of this thesis consists of a set of interconnected gestural artificial language experiments. These were designed to fit into different sections of the

---

14 The signed language data in this Table was collated by me based on available data from Coons (2022). The included data is limited to those languages where a clear order could be defined for all dependent types. Unfortunately this means there is an extremely small sample size for signed languages in Table 1.6 and Table 1.5. With overlapping data from more signed languages, the parallel between the pattern in spoken languages and signed languages could be explored in more detail.
Table 1.5: Typological counts of the order of nouns with adjectives and numerals in spoken (spo.) and signed (sig.) languages (Coons, 2022; Dryer, 2013a, 2013d).

<table>
<thead>
<tr>
<th>Order</th>
<th>N (Spo.)</th>
<th>N (Sig.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-Num/N-Adj</td>
<td>510</td>
<td>9</td>
</tr>
<tr>
<td>Num-N/Adj-N</td>
<td>251</td>
<td>6</td>
</tr>
<tr>
<td>Num-N/N-Adj</td>
<td>168</td>
<td>3</td>
</tr>
<tr>
<td>N-Num/Adj-N</td>
<td>37</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1.6: Typological counts of the order of nouns with adjectives and genitives in spoken (spo.) and signed (sig.) languages (Coons, 2022; Dryer, 2013a, 2013c).

<table>
<thead>
<tr>
<th>Order</th>
<th>N (Spo.)</th>
<th>N (Sig.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-Gen/N-Adj</td>
<td>342</td>
<td>-</td>
</tr>
<tr>
<td>Gen-N/Adj-N</td>
<td>232</td>
<td>8</td>
</tr>
<tr>
<td>Gen-N/N-Adj</td>
<td>342</td>
<td>9</td>
</tr>
<tr>
<td>N-Gen/Adj-N</td>
<td>65</td>
<td>-</td>
</tr>
</tbody>
</table>

proposed Scale of Innovation I outlined above in Section 1.5. The overarching goal of these experiments is to examine where along the Scale of Innovation that we see participants’ behaviour being influenced by category-specific and system-wide cognitive biases respectively. Note that each of the three main content chapters are directly preceded by a short preface that situates the chapter within the wider context of the thesis. Additionally, each main content chapter is followed by an associated bibliography and appendix. The full thesis bibliography, encompassing works cited across all chapters, can be found at the end of the thesis.

Chapter 2 includes a set of two online experiments, conducted with two main goals in mind; establishing the existence of category-specific biases for the two noun phrase dependents, and examining if these biases also influenced the learnability of different noun phrase word orders. The result of the first, gestural experiment show that these biases are present and observable in the behaviour of participants in a context where specific noun phrase meanings are expressed and judged in the absence of a wider linguistic system. The second experiment constitutes a gestural artificial language learning experiment using a regularisation design to examine if the preferred orders from the first experiment also imparted a learning advantage for artificial languages with just one meaning type (either adjectives or genitives). The results of the second experiment showed no indication that word orders favoured by the category-specific biases in the first experiment were learned better, or regularised more readily, than alternative orders.

Chapter 3 builds on the last experiment in Chapter 2, continuing to explore the possibility that favoured noun phrase word orders could be learned better, but in a slightly different context. In this gestural artificial language learning experiment the language participants were trained on was more complex, consisting of both possessive and descriptive noun phrases. In this experiment the category-specific biases were thus competing with cross-category harmony, rather than with within-category harmony as in Experiment 2 in Chapter 2. With this design I explored if
category-specific biases modulated participants’ tendency for cross-category word order harmony. The results show no modulating effect of harmonisation based on category-specific biases, but did find evidence that participants employed several different strategies for reducing unpredictable variation in their output, which are in line with system-wide biases favouring both harmony and regularity.

Chapter 4 introduces a set of two experiments, designed in parallel to allow for cross-experiment examination of harmonisation and regularisation behaviour in different contexts. Specifically, these experiments target the middle-region of the Scale of Innovation introduced above in section 1.5, to examine if contexts that involve a mixture of learning and linguistic innovation constitute the locus of competition between system-wide and category-specific biases. The main manipulation across the experiments is in how much of that language system that participants are exposed to during training. In the first experiment participants are trained and tested on exactly the same items, meaning hardly any linguistic innovation is required during the testing phase. In the second experiment participants are trained on one item from each dependent type (one adjective and one genitive) and tested on meanings including the remaining item from each of these categories. The task at test thus requires generalisation (within category extension) from the meanings participants saw in training, to the novel meanings they see at test. The main prediction was that we would see more influence of category-specific biases in the generalisation than in the learning experiment (although we also acknowledged the potential for a different set of predictions). Results here show a slightly higher tendency for participants to be influenced by category-specific biases in the second experiment, suggesting that the kind of innovation that is involved in a generalisation task is enough to (at least partially) activate the category-specific biases for these dependent types. In addition, orders favoured by category-specific preferences were also learned slightly better, echoing previous results showing how category-specific biases can come to be reflected in learning behaviour.

Finally, Chapter 5 summarises the overall results of the thesis and reflects on the impact of the research in light of these new findings. It examines how the thesis has contributed to our understanding of the way cognitive biases affect typological structure, and expands on what these results mean for our conceptualisation of the trajectory of language evolution. Additionally, it highlights the implications of specific design choices and theoretical decisions that were made throughout this thesis project. Finally, it includes a discussion of potential future directions that can be pursued based on these results, and how these would further enrich our knowledge of the way cognitive biases active within individuals
contribute to population-wide phenomena across linguistic domains.
Chapter 2

With or without a system: How category-specific and system-wide cognitive biases shape word order

2.1 Preface: Chapter 2

While Chapter 1 established the typological background for the existence of category-specific biases affecting adjectives and genitives, this needs to be supplemented by experimental data showing that there is behavioural evidence for the existence of these ordering preferences when individuals interact with linguistic data. The first experiment in the following chapter does just that. It sets out to establish if these biases can be observed in an experimental task where individual gestural expressions of adjective and genitive meanings are presented to participants without reference to how these expressions fit into a wider language system. This task is thus situated within the improvising region of the Scale of Innovation, but with a judgement-based task rather than a task involving production-based improvisation. The choice of a judgement-based task over a production-based task was motivated by both practical and theoretical implications. Firstly, since data collection for the research in this thesis was mainly carried out during the COVID-19 pandemic all research had to be conducted online, and collecting audio data online requires recording individuals in their own home using their own audio equipment. This presents challenges both for maintaining participants’ privacy, increases the potential for extensive data-loss and requires a much more time-consuming data analysis procedures than the judgement-based task we use here. Secondly, a judgement-based task is a harder test for these category-specific biases than a production-based task. This is due to the fact that participants are
given the option between the theoretically favoured and theoretically disfavoured variants (here gesture orders), whereas in a production-based task they may never even consider the option of the disfavoured variant. I return to this point in the Discussion chapter, section 5.2.

Note that the main text in this chapter does not contain clear references to the Scale of Innovation. This is because the theory behind the scale was still under development when data collection for these experiments took place. The first experiment finds behavioural evidence for the postnominal adjective and prenominal genitive preference which supplement the patterns we have already observed in the typological data of both spoken and signed languages.

On the back of this result I examine if these same biases would also influence participant’s learning behaviour, such that language systems which order adjectives postnominally and genitives prenominally would be easier for participants to learn. Experiment 2 explores this using a gestural artificial language learning methodology with variable training data, meaning it relies on participants’ regularisation tendency to reveal differences in word order preferences. This context sits much further to the left along the Scale of Innovation, firmly in the regularising region of the scale, and thus requires much less innovation than Experiment 1. If participants in Experiment 2 were to learn bias-congruent systems more easily, then this would suggest that the category-specific biases for postnominal adjectives and prenominal genitives are strong enough to compete with the system-wide preference for harmony, even during tasks that require very little innovation. The motivation behind examining this possibility came from two main sources, firstly the observation outlined in Chapter 1, that biases which are active in learning will have iterative effects on language structure with each new generation of learners, and results observed by Motamedi et al. (2022) and Do et al. (2022), which suggest that some category-specific biases do indeed have this type of effect. If the category-specific biases for adjectives and genitives were found to influence learning here it would help explain why we see their effects so clearly in typology. The results of Experiment 2 show no learning advantages or higher tendency for regularisation for language systems with orders that are favoured by category-specific biases, but does see evidence of both a bias for harmony and a tendency to reduce unpredictable variation.

In the appendix for this chapter you will also find more detailed explanations for why I chose the specific meanings that I use throughout this thesis, and how the gestural stimuli was created. Furthermore, there is information on how I performed the Monte Carlo simulations that are used to validate the reliability of changes in entropy and mutual information that are reported in the results
section for Experiment 2. Finally, there are several summaries of demographic and translation data based on the questions participants answered at the end of both experiments.
2.2 Holtz et al. (under revision): author contributions

The following paper was submitted to Cognitive Science in May 2023. The paper was co-authored with my two supervisors, Jennifer Culbertson and Simon Kirby. The development of the hypotheses, experimental design, and analysis took place during supervision meetings where all three authors were present and contributing. Both co-authors gave feedback during the write-up process. The manuscript has been reformatted to adhere to thesis formatting guidelines. Discrepancies between this paper and any future published version may occur as a result of the review process.
2.3 Abstract

There are certain recurrent features of language that characterise the way whole language systems are structured, and others that target specific categories of items within those wider systems. For example, languages tend to exhibit harmonic (i.e. consistent) ordering between heads and dependents, making it a kind of system-wide regularity. While this tendency is generally robust, some categories of linguistic items deviate from this trend. We examine one such case of non-harmony, namely the order of the noun with respect to two dependents – adjectives and genitives. Using two silent gesture experiments, this study tests the hypothesis that category-specific cognitive biases favour postnominal adjective order and prenominal genitive order. Further, we show that the influence of these biases is revealed in contexts where no conventionalised system is in place. When a system is in place, participants learn that system, and category-specific biases do not impact their learning. Our results suggest that different types of linguistic contexts reveal the influence of separate types of cognitive biases, such that some are active during learning and others are active during language creation.

**Keywords:** cognitive biases; typology; silent gesture; regularisation; word order
2.4 Introduction

Typological research has revealed a variety of regularities in how languages order meaningful elements such as affixes, words, and phrases (Coons, 2022; Dryer, 1992; Greenberg, 1963; J. A. Hawkins, 1990). The underlying causes of these typological patterns remain an open question in linguistics, but there are several proposed explanations. These include innate constraints on language structure (Chomsky, 1993; Travis, 1984), lineage-specific trends in language history (Dunn et al., 2011; Piantadosi & Gibson, 2014), common processes of language change (Bybee, 2008; Collins, 2019), and cognitive biases which tend to (dis)favor certain linguistic structures (Culbertson et al., 2012; Finley, 2015, 2018; Martin et al., 2020; Saldana et al., 2021; Wilson, 2006). Research examining the latter type of explanation has explored the role that cognitive biases play in a variety of linguistic phenomena, such as the type of phonological rules present in language (Finley, 2015; Martin & White, 2019; Wilson, 2006), the order of derivational and inflectional morphemes (Saldana et al., 2021), the prevalence of certain basic word order patterns (Goldin-Meadow et al., 2008; Hall et al., 2013), and the tendency for languages to exhibit harmonic (i.e. consistent) order between heads and dependents (Christiansen, 2000; Culbertson, Franck, et al., 2020; Culbertson et al., 2012).

These studies use controlled experiments to test whether linguistic patterns that are more common across languages are preferred by participants. In some cases, the patterns tested are present at the level of an individual word, phrase, or utterance. For example, Martin et al. (2020) find a preference for placing adjectives closest to nouns in a noun phrase with multiple nominal dependents. Other patterns describe features of a language that hold across phrases or utterance. For example Culbertson, Franck, et al. (2020) show that learners prefer placing dependents consistently before or after nouns, across phrases in a language. Here, we will refer to the former type of preference as category-specific, and the second as system-wide. Biases that target either of these levels of analysis have been found in multiple linguistic domains. Below, we review studies on nominal and basic word order which illustrate how category-specific and system-wide biases in some cases conflict. We will use such a case in order to explore how, and under what circumstances these different types of biases come to influence language typology.
2.4.1 Category-specific and system-wide biases

Evidence for category-specific and system-wide biases in word order come from experiments using artificial language learning (ALL) and silent gesture paradigms. For example, Culbertson and Adger (2014) and Martin et al. (2020) use an ALL paradigm in which participants are trained on a novel language in which nouns are modified by either an adjective, a numeral, or a demonstrative, but the relative order of these dependents is withheld. At test participants tend to infer orders where the adjective is closest to the noun, followed by the numeral, and with the demonstrative furthest from the noun (regardless of whether dependents are prenominal or postnominal). These orders are argued to be preferred because they transparently reflect, or are homomorphic to a single representation in which adjectives are grouped most closely with the noun, and demonstratives furthest (see e.g. Culbertson & Adger, 2014). Similar results have also been obtained in silent gesture studies where hearing non-signing participants have to create new ways to express complex noun phrase meanings using just their hands (Culbertson, Schouwstra, & Kirby, 2020). The orders that individuals favour in these experiments – the so-called homomorphic orders – are also the orders we find most commonly in typological data based on spoken languages (see Figure 2.1 for visualisation of different orders across spoken languages), suggesting that biases active at the level of the individual might be a driving force behind some cross-linguistic regularities. This kind of bias falls under our definition of category-specific, since it targets specific, individual instances of a given type of phrase.

Similarly, category-specific word order preferences exist for basic word order, i.e., the relative order of subject, object, and verb in declarative sentences. While there seems to be a general bias in favour of S(ubject)O(bject)V(erb) to express basic transitive events (Goldin-Meadow et al., 2008; Langus & Nespor, 2010)—a category-specific bias in and of itself—some studies have argued that even more specific features of the referents involved in the event (e.g. animacy or salience of the referents) (Gibson et al., 2013; Hall et al., 2013; Kirton et al., 2021; Meir et al., 2017), or the semantics of the verb itself (Schouwstra & de Swart, 2014) can reveal more fine-grained, category-specific preferences for basic word order. Schouwstra and de Swart (2014) found that both Dutch- and Turkish-speaking participants had the same basic ordering preferences, conditioned by the type of event that they were gesturing. When they were tasked with conveying extensional meanings like *throw* or *carry* they tended to produce SOV order, but when conveying intensional events like *hear* or *dream of* they tended to use SVO order. Similar
results have also been observed with English speakers (Motamedi et al., 2022).

A key difference between the two examples of category-specific biases we have exemplified is that, while the typological tendency towards homomorphic structure is robust, there are few languages that seem to have a productive distinction in word order based on event type (so far, it has been found in Brazilian Sign Language; Napoli et al. (2017), and in Nicaraguan Sign Language; Flaherty et al. (2018)). This could be due to another, more general, tendency in favour of having a consistent basic word order. Using one word order consistently reduces variability across the language system, and there is experimental evidence showing that people tend to reduce variability, especially if the task involves linguistic stimuli (Culbertson et al., 2012; Ferdinand et al., 2019; Samara et al., 2017; Smith & Wonnacott, 2010; Smith et al., 2017). The preference to reduce variation in word order could lead to the loss of patterns arising from category-specific biases, such as conditioning based on event type. Successive generations of learners may impose the more systematic use of one word order, with the other variant gradually being lost. In other words, there may be a system-wide bias for word order consistency that competes with category-specific biases for word order to make the language system more regular.

A system-wide preference for consistent word order in a language can also be
observed at a more abstract level in the typological tendency for languages to exhibit harmonic order between heads and dependents (Culbertson & Kirby, 2016). The tendency for languages to be harmonic can be seen both across and within phrase types, such that individual languages tend to be either head-final or head-initial across phrases, and dependents of the same head tend to fall on the same side of that head (Dryer, 1992; Greenberg, 1963; J. A. Hawkins, 1990). While there are suggested historical explanations for this pattern (Dunn et al., 2011; Piantadosi & Gibson, 2014), experimental research using ALL has also shown that participants tend to learn harmonic word order better than non-harmonic word order. For example, studies employing a regularisation paradigm, where participants are trained on variable word order, find regularisation of variable harmonic systems more readily than non-harmonic ones (Culbertson, Franck, et al., 2020; Culbertson et al., 2012). This suggests that there may be an active bias, at the level of the individual, which contributes to the robustness of this typological pattern. Such a bias would crucially be active at the level of the linguistic system, rather than at the level of individual linguistic items within that system, since harmony constitutes a structured relation between items in a broader linguistic system. As in the case of basic word order and event type conditioning, a system-wide bias for harmony may also be in competition with category-specific biases.

Indeed, there are clear exceptions to word order harmony. For example, although noun phrase dependents tend to exhibit harmony, certain dependent types are more likely to stand out as exceptions. In particular, adjectives (the red house) are more likely to be postnominal, and genitives (the child’s toy) are more likely to be prenominal, regardless of the order of other dependents (see Table 2.1 and Table 2.2, data based on spoken language data in WALS (Dryer, 2013a, 2013b)). The impact that this has on harmony can be seen by looking at the number of spoken languages exhibiting harmonic vs. non-harmonic orders of these two elements relative to the noun (see Table 2.3, data based on spoken language data in WALS (Dryer, 2013a, 2013b)). This table shows that non-harmonic order where the genitive precedes the noun and the adjective follows it is just as common as the postnominal harmonic order. The prenominal harmonic order and and the non-harmonic order with prenominal adjectives but postnominal genitives are both much less common. A similar pattern is also observed in typological data based on sign languages, where most languages exhibit postnominal ordering for adjectives and prenominal ordering for genitives (Coons, 2022). This deviation from the harmonic pattern suggests that there may be two category-specific ordering biases—one for postnominal adjectives and one for prenominal genitives—that
compete with a system-wide bias for harmony.

Table 2.1: Order of adjectives in relation to nouns showing postnominal preference.

<table>
<thead>
<tr>
<th>Order</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun-Adjective</td>
<td>879</td>
</tr>
<tr>
<td>Adjective-Noun</td>
<td>373</td>
</tr>
<tr>
<td>Other</td>
<td>110</td>
</tr>
</tbody>
</table>

Table 2.2: Order of genitives in relation to nouns showing prenominal preference.

<table>
<thead>
<tr>
<th>Order</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun-Genitive</td>
<td>468</td>
</tr>
<tr>
<td>Genitive-Noun</td>
<td>685</td>
</tr>
<tr>
<td>Other</td>
<td>96</td>
</tr>
</tbody>
</table>

Table 2.3: Order of adjectives and genitives in relation to nouns showing that the two most common orders is the postnominal harmonic order and the Gen-N-Adj order.

<table>
<thead>
<tr>
<th>Order</th>
<th>Noun-Adjective</th>
<th>Adjective-Noun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun-Genitive</td>
<td>342</td>
<td>65</td>
</tr>
<tr>
<td>Genitive-Noun</td>
<td>342</td>
<td>232</td>
</tr>
</tbody>
</table>

How might such category-specific cognitive biases come to influence typology and compete with system-wide biases like harmony? While in principle, both types of biases might be driven by the same mechanisms, and emerge under the same conditions, here we explore the possibility that category-specific biases emerge most strongly during improvisation, where no system that could activate system-wide biases is in place.

2.4.2 Different biases in different contexts

The research reviewed above suggests an intriguing pattern: evidence for system-wide and category-specific biases appears in different contexts. For example, most experiments revealing a bias for word order harmony involve participants learning a language system and being asked to reproduce it (Christiansen, 2000;
Culbertson, Franck, et al., 2020; Culbertson et al., 2012). In a task where participants have to improvise the linguistic structure themselves, producing gestures for complex noun phrases without any model/input, harmonic orders are not preferred (Culbertson, Schouwstra, & Kirby, 2020). Conversely, it is in precisely these improvisational experimental contexts that category-specific preferences have been found. The bias for homomorphism was found in studies where participants either improvise in the absence of any conventionalised system, or innovate the relevant part of the system (Culbertson, Schouwstra, & Kirby, 2020; Martin et al., 2019, 2020). Similarly, preferences for basic word order patterns specific to particular categories of verbs or event types have emerged under these same conditions (Motamedi et al., 2022; Schouwstra & de Swart, 2014).

These findings suggest the possibility that system-wide and category-specific biases emerge via distinct mechanisms, or at least, in different linguistic contexts. Category-specific biases may influence behaviour most strongly in contexts requiring improvisation, when there is no language system already in place (or relatively little evidence for one); system-wide biases may influence language most during learning, where the different parts of the system are all in play. If this is true, then category-specific biases may have a relatively limited window in which to influence the evolution of language structure, e.g., early during language emergence, or when a completely novel structure or combination must be improvised. System-wide preferences, on the other hand, would influence language continuously in contexts of language learning, exerting pressure anew on languages at each generation in a way which can amplify the effect of weak individual biases (Reali & Griffiths, 2009; Thompson et al., 2016).

However, some category-specific biases appear to be surprisingly persistent—like the typological tendency for adjectives to be postnominal, and genitives prenominal. This suggests that at least some category-specific biases might also influence learning. In other words, these biases may emerge not only in improvisation, but may also make a linguistic system which aligns with them easier to learn. Few studies have directly investigated this possibility, but recent work has found some evidence that systems in which basic word order aligns with category-specific biases for event type are indeed easier to learn (Motamedi et al., 2022).

The experiments we report here investigate the possibility that category-specific biases affect both improvisation and learning in the domain of noun phrase

\footnote{Wang et al. (2023) find that participants trained on the head-dependent order in the verb phrase (e.g., VO or OV) generalise this order to the adposition phrase (e.g., Preposition-Noun or Noun-Postposition respectively). However, it may be that there is enough of a system in place here to trigger the relevant bias.}
word order. In a first experiment, we provide the first behavioural evidence of a preference for postnominal adjectives and prenominal genitives in a silent gesture task where there is no conventionalised linguistic system in place. We then test whether these preferences continue to influence participants' behaviour in a silent gesture learning experiment. Here, the testing materials are identical, but are preceded by a stage in which participants are exposed to evidence for a conventionalised gesture order system they must learn.

2.5 Experiment 1

In this first experiment, we use a silent gesture judgement task to test participants' ordering preferences for genitives and adjectives in the absence of evidence of a wider, conventionalised, linguistic system. The experiment uses a between-subjects design, manipulating dependent meaning: descriptive (adjective) or possessive (genitive). Following the method used in Motamedi et al. (2022), participants were given a single trial in which they were asked to choose between two gesture videos. Here, one video uses a prenominal gesture order for the dependent and the other a postnominal gesture order for the dependent meaning.\footnote{This experiment was granted ethical approval by the PPLS Research Ethics Committee. The study was also preregistered https://shorturl.at/mqEY8 and all associated materials and code can be viewed at https://shorturl.at/dJQU2.} We predict that participants will exhibit a preference for postnominal order in the adjective condition, but prenominal order in the genitive condition. If these preferences are found, this would support the notion that the ordering of these dependent two types are subject to category-specific biases.

2.5.1 Methods

Materials

The experiment was developed using the JavaScript library jsPsych (de Leeuw, 2015) and ran in participants’ web browsers. Participants saw a collection of grayscale digital drawings showing either instances of item ownership (genitive condition, e.g. vampire’s hat) or items with different patterns (adjective condition, e.g. striped cup). The set of images consisted of every possible combination of the two genitives (possessors) vampire and cyclops, the two adjectives spotted and striped as well as four nouns hat, scarf, cup, and book. The images were created in Inkscape and, in total, there were 16 possible images each representing a different meaning (see Figure 2.2 for sample images from each condition).
For each image there were two gesture videos, making a total of 32 videos. The videos showed a model gesturer producing two gestures in sequence, one representing the head noun and one representing the dependent, either an adjective or genitive. The videos differed only in the order of these two gestures – in one the head noun was the first gesture, in the other it was the last. Each phrase component was denoted using a gesture made with both hands and the videos ended with both hands in a neutral position. The videos were all 4,389 milliseconds long and matched so that the beginning and end of each component gesture was synchronised across each pair of videos.

Figure 2.2: These samples shows a subset of the total possible set of stimuli images.

**Procedure**

Participants were randomly assigned to either the adjective condition or the genitive condition. They were instructed that the study was about ‘how to describe items in a sign language’ if they were in the adjective condition, or ‘how you express ownership in a sign language’, if they were in the genitive condition. Prior to the main testing trial, participants were shown a sample 2x2 grid of images containing the kinds of images that they would be shown in the test trial and these images exhibited contrasts along both the head and dependent dimensions (see Figure 2.3a).

The instructions for this familiarisation trial necessarily included reference to an ‘item’ and either a ‘pattern’ or ‘owner’. The order in which the head noun referent ‘item’ and the two dependent referents (‘pattern’ and ‘owner’) appeared
in the instructions was randomised between participants. This was done to avoid the possibility that the order between these elements in the instructions biased participants in favour of a certain gesture order in the test trial. The alternative instructions that participants saw included either ‘Across the images both the patterns and the items vary’ or ‘Across the images both the items and the patterns vary.’ in the adjective condition and ‘Across the images both the owners and the items vary’ or ‘Across the images both the items and the owners vary’ in the genitive condition. We examine the effect of this ordering as part of our pre-registered exploratory analysis in the results section.

After this pre-test trial, participants were instructed that they would see the same kind of 2x2 grid but with one image highlighted in red. They were told that two videos would appear next to the image grid and that these represented two ways to express ‘ownership of the item’ in the highlighted image (in the genitive condition) or two ways that the highlighted ‘item could be described’ (in the adjective condition) in a made up sign language. Their task was to choose the gesture video which they thought best conveyed the meaning of the highlighted image. The images remained on the screen, with the videos looping next to them, until participants chose one gesture video by clicking on it (see Figure 2.3b for an example trial).

Follow this single test trial, a second trial asked participants to drag the point of a slider to indicate how strong their preference was for the gesture order they chose in the previous forced-choice trial. The target image from the forced-choice trial was displayed above the slider and the two videos looped on either side of the slider and were labelled ‘A’ (for the left video) and ‘B’ (for the right video). The location of the prenominal and postnominal video was randomised per participant. To submit a response, participants had to drag the slider point from the middle towards one of the videos. The slider was marked with ‘weakly prefer video A/B’ and ‘strongly prefer video A/B’ on either side of the mid-point (see Figure 2.3c for example of slider trial). Following this, participants were shown the video they had chosen in the forced-choice trial and were asked to translate the meaning of the gesture video into English by typing in a response. After this trial participants were asked to provide a motivation for why they chose the specific video in a free type trial.\(^3\) Finally, participants responded to two short demographics questions. One asking them if they knew a sign language (used for preregistered exclusions) and another asking them to note which spoken

---

\(^3\)Translation trials were mainly used to ensure that the visual stimuli conveyed the intended meanings. Overviews of the translation data can be found in the R scripts in the project OSF repository. Motivations given for participants’ choices can also be found in the survey data.
languages they knew and at what level of proficiency (on a scale of 1-10, where 10 was native-like proficiency). A total of 384 participants were recruited via the online crowdsourcing platform Prolific. Using the built-in Prolific prescreening options, we restricted participation to those who reported English as their first language, had at least a 95% proficiency in English, and lived in the United States. The most commonly reported L2s for participants in both conditions were French, Spanish and German. All of these languages are reported as mainly having postnominal order for genitive meanings in WALS (Dryer, 2013b). French and Spanish are reported as mainly having postnominal adjectives (although many of the most common adjectives in these languages also appear prenominally) and German is reported as mostly having prenominal adjectives (Dryer, 2013a). Full list of L2s reported by participants in Experiment 1 can be found in the OSF repository.

Figure 2.3: Illustrative examples of the main trials in Experiment 1.

Participants

A total of 384 participants were recruited via the online crowdsourcing platform Prolific. Using the built-in Prolific prescreening options, we restricted participation to those who reported English as their first language, had at least a 95% proficiency in English, and lived in the United States. The most commonly reported L2s for participants in both conditions were French, Spanish and German. All of these languages are reported as mainly having postnominal order for genitive meanings in WALS (Dryer, 2013b). French and Spanish are reported as mainly having postnominal adjectives (although many of the most common adjectives in these languages also appear prenominally) and German is reported as mostly having prenominal adjectives (Dryer, 2013a). Full list of L2s reported by participants in Experiment 1 can be found in the OSF repository.
previous task approval rate, and had not completed any of our previous experiments or pilots. Participants were paid the equivalent of £8.91 per hour. We excluded 8 participants who stated that they were proficient in a sign language. A further 56 participants were excluded as they responded too quickly to the forced-choice trial (< 9.678 milliseconds, combined time for both videos, meaning they had not watched both videos before making their choice; N=18), did not indicate a preference for the same gesture video across both the forced-choice and slider trial (N=24), or both (N=14). After these (preregistered) exclusions there were 160 participants in each condition.

2.5.2 Coding

The forced-choice trial responses were coded using a binary variable, predicted order, with 1 for the predicted order (prenominal order in the genitive condition, postnominal in the adjective condition) and 0 for the alternative order (postnominal in the genitive condition and prenominal in the adjective condition). The slider trial responses were transformed to account for the fact that values close to 0 represented a strong preference for the video on the left, and a value very close to 100 represented a preference for the video on the right. To make these preferences comparable, independently of video placement, all values under 50 were converted to their corresponding value above 50 (e.g. 2 becoming 98).

2.5.3 Results

Based on the typological data, we made two main predictions for Experiment 1: (i) participants will prefer the postnominal gesture order when the gestures represented a descriptive meaning, (ii) participants will prefer the prenominal gesture order when these represented a possessive meaning. We also made an additional prediction based on the typological data, where the asymmetry in the prenominal vs postnominal order for adjectives and genitives is such that the postnominal adjective preference appears (numerically) stronger than the prenominal genitive preference (Dryer, 2013a, 2013b). Therefore we predicted: (iii) the postnominal preference for descriptive meanings will be stronger than the prenominal preference for possessive meanings.

Main analysis

To evaluate our first two predictions, we first examined the extent to which participants chose the predicted order in forced-choice trials across the two conditions.
As shown in Figure 2.4, participants’ choices closely match what is observed in the typological data for spoken languages (Dryer, 2013a, 2013b). According to our preregistered analysis plan, the data were analysed using mixed effects logistic regression models implemented using the lme4 package (Bates et al., 2015) in R (R Core Team, 2022). Results from two intercept-only models (one per condition), with predicted order as the outcome variable, indicated that participants chose the predicted order for their respective conditions at rates significantly above chance (genitive condition: $\beta = 0.56, SE = 0.16, z = 3.43, p < 0.001$, adjective condition: $\beta = 0.51, SE = 0.16, z = 3.02, p < 0.01$). These results support our first two predictions.

Figure 2.4: Proportion of postnominal orders per dependent meaning based on the typological data (left facet) and selections of participants in the forced-choice trials (right facet). In both the typological and the experimental data, postnominal order is preferred for adjectives and prenominal order is preferred for genitives.

We can also test whether participants who chose the predicted order on the forced-choice trial also showed a stronger preference for these orders (i.e., gave

---

5All data were analysed in this way unless otherwise noted.
6A post-hoc analysis examining whether participants in the adjective condition who indicated having knowledge of a language with postnominal adjective order were also more likely to choose a postnominal gesture order in the forced-choice trial (predicted order $\sim$ postnominal language) revealed no such preference $\beta = 0.22, SE = 0.34, z = 0.77, p = 0.51$.
7Note that the model for the adjective condition included a random effect for items (i.e., the specific target image participants were exposed to). The model with a random effect for item did not converge in the genitive condition, therefore a logistic regression model with no random effect is reported here for genitives.
them a higher rating) in the slider task, compared to those who chose the alternative order. Figure 2.5 shows the number of participants in each condition who indicated a given preference strength on the slider trials. To analyse these data we ran two linear models, one for each condition. The outcome variable was the transformed rating values, with a fixed effect of predicted order. While Figure 2.5 suggests that, when plotting counts of participants who chose a specific preference strength, more participants gave the highest rating to predicted orders in each condition, neither model reached significance (genitive condition: $\beta = 3.89$, $SE = 2.14$, $t = 1.81$, $p = 0.071$, adjective condition: $\beta = 1.06$, $SE = 2.07$, $t = 0.51$, $p = 0.71$).\(^8\) To summarise, participants were more likely to choose the predicted order on forced-choice trials. However, there is no evidence that preference ratings in the slider task were stronger for participants who chose the predicted order, compared to those who did not.

To evaluate our third prediction, that the postnominal adjective preference would be stronger than the prenominal genitive preference, we again analysed both the forced-choice and slider data. We analysed the forced-choice data using a logistic regression model with predicted order as the outcome variable and condition as a fixed effect (adjective condition acted as baseline). This model revealed no difference between the adjective and genitive conditions ($\beta = 0.05$, $SE = 0.23$, $z = 0.23$, $p = 0.72$).

We also evaluated this prediction for the slider data, using a linear model with transformed rating as the outcome variable, and predicted order, condition (adjective condition acted as baseline), and their interaction as fixed effects. The results revealed a significant negative coefficient for condition ($\beta = -4.93$, $SE = 2.36$, $t = -2.09$, $p = 0.04$) indicating that preference ratings were, overall, slightly lower in the genitive condition than in the adjective condition. However, no significant interaction was found. To summarise, neither the forced-choice data nor the slider data provide evidence for a stronger preference for the predicted order in the adjective condition compared to the genitive condition, thus not mirroring the numerical difference between prenominal genitives and postnominal adjectives found in the typological data.

**Exploratory analysis**

In addition to testing our three preregistered main predictions, we also conducted an exploratory analysis. This was done to rule out the possibility that the order

---

\(^8\)Neither model converged with random intercepts for item and so this effect was removed from the model.
of words in the instructions for the familiarisation trial influenced the gesture order participants chose in the forced-choice task (see section 2.5.1 for details about word order in instructions). We ran two logistic regression models, one per condition, with the binary outcome variables *prenominal* (genitive condition) and *postnominal* (adjective condition) with the two-level fixed effect of *instruction order* (prenominal or postnominal; postnominal as baseline). Neither of these models reached significance (genitive model: $\beta = -0.25$, $SE = 0.33$, $z = -0.75$, $p = 0.46$, adjective model: $\beta = -0.12$, $SE = 0.33$, $z = -0.37$, $p = 0.71$), therefore there is no evidence that the order of the words in the instructions determined the choice participants made in the experimental forced-choice trial.

### 2.5.4 Experiment 1: Discussion

In Experiment 1 participants were tasked with choosing (and rating) their preference for a gesture order to express either an adjective or genitive meaning. They did this in the absence of any evidence about the wider linguistic system, and for
only a single exemplar meaning. The results confirm our main predictions: the orders participants preferred for descriptive and possessive meanings in a silent gesture preference task match the most common orders we see for adjectives and genitives in both spoken and sign language typology (Coons, 2022; Dryer, 2013a, 2013b). This suggests that the typology may reflect category-specific biases for these two types of dependent meanings. These bias may also explain the absence of harmony between these two dependent types: category-specific ordering preferences may work to keep these dependents split across the head noun.

While we have found clear evidence for these two category-specific biases, we failed to find any evidence that one was stronger than the other. There was no overall difference in the likelihood of choosing the predicted order across our two conditions, and the preference ratings were not stronger for postnominal adjectives compared to prenominal genitives.

It could be that this particular asymmetry simply reflects some other mechanism—like accidental facts about language history—since of course simply counting the numbers of languages that have one pattern versus another does not control for genetic or areal relationships among languages. However, it is also possible that the lack of difference in the strength of participants’ preferences could reflect some influence from their native language. English has variation in the order of genitives, and even some variation in the order of adjectives, however for the types of meaning used in this experiment, prenominal order is preferred for both (e.g., vampire’s hat, spotted hat). A prenominal preference coming from English experience might therefore strengthen the more general preference for prenominal genitives, but weaken the more general preference for postnominal adjectives. This would reduce the difference between the two dependent types in the experiment. However, given that our exploratory analysis of the impact that knowledge of a language with postnominal adjective order had no effect on adjective order choice it is unclear how much of an impact that spoken language transfer had on participants’ order preferences.

It is also possible that the lack of difference between the dependent types is due to a fundamental difference in how these category-specific preferences operate in the vocal/auditory modalities versus the manual/visual modalities. Previous research has shown that some ordering preferences observed in silent gesture studies may be influenced by modality-specific constraints (Napoli & Sutton-Spence, 2014; Struhl et al., 2017). Similar effects could be at work here, strengthening the prenominal genitive preference. For example, in the gestures denoting the possessor used in this experiment, the body of the gesturer is used to inhabit the role of the animate referent (either the vampire’s body or the cyclops’ body).
Studies on sign languages and silent gesture have found that signs which make use of the body in this way are often linearised earlier in production, which, in this case, would create a stronger prenominal genitive preference (Meir et al., 2017). This type of modality-specific effect could explain why the typological data from sign languages show the same directionality of ordering preferences (Coons, 2022), but a (numerically) stronger preference for prenominal genitives compared to postnominal adjectives.

Regardless of the lack of difference in the strength of the two preferences, we have found here clear evidence for category-specific preferences influencing order in the absence of a linguistic system. In other words, in this task, where participants received no evidence of a conventionalised linguistic system, postnominal adjectives and prenominal genitives were clearly the preferred orders for these dependents, in accordance with typological evidence. In the next experiment we ask whether these same orders are also easier to learn.

2.6 Experiment 2

Experiment 2 tested whether the ordering preferences observed in the Experiment 1 also influence how participants learn the order of adjectives and genitives with respect to nouns. In principle, one could simply ask whether fixed ordering systems that use one of the preferred orders (e.g., postnominal adjectives or prenominal genitives) are easier to learn. However, there is evidence that in simple artificial language learning experiments, adults very easily learn and reproduce even very rare noun phrase orders (Culbertson & Newport, 2017). Therefore, we follow others in using a regularisation design, in which participants are exposed to a system with some unpredictable variation. Specifically here, multiple orders are possible (e.g., both prenominal and postnominal adjectives), but one order is more common. This design capitalises on the fact that learners tend to regularise rather than reproduce unpredictable variation (Ferdinand et al., 2019; Hudson Kam & Newport, 2009; Smith & Wonnacott, 2010), but regularisation is more likely when the majority order is preferred due to some other factor (Culbertson & Newport, 2015; Culbertson et al., 2012).

We test whether participants are more likely to learn and regularise the majority order they are exposed to when this order aligns with the category-specific biases identified in Experiment 1—i.e., postnominal adjectives or prenominal genitives. Finding evidence that the preferences from Experiment 1 continue to influence participants’ behaviour when they are tasked with learning an existing
language system could help explain why these two dependent types resist the tendency for harmony. More generally, if category-specific preferences are active not only during language emergence, but also during language learning, then they have more opportunity to influence language structure.

Experiment 2 was a between-subjects silent gesture design, similar to (Motamedi et al., 2021). There were four conditions, created by crossing the two variables of interest, namely dependent type (either adjective or genitive), and what we will call the naturalness of the majority order that participants were trained on (either natural or unnatural), where natural orders were those which aligned with the category-specific preferences found in experiment 1. The conditions were called natural adjective, unnatural adjective, natural genitive, and unnatural genitive. Participants were first trained on example gestures in each condition, as shown in table 2.4. They then completed the same type of forced-choice judgement task as in Experiment 1.\textsuperscript{9}

Table 2.4: Percentage of prenominal and postnominal gesture orders in input per condition

<table>
<thead>
<tr>
<th>Condition</th>
<th>Prenominal</th>
<th>Postnominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Adjective</td>
<td>25%</td>
<td>75%</td>
</tr>
<tr>
<td>Unnatural Adjective</td>
<td>75%</td>
<td>25%</td>
</tr>
<tr>
<td>Natural Genitive</td>
<td>75%</td>
<td>25%</td>
</tr>
<tr>
<td>Unnatural Genitive</td>
<td>25%</td>
<td>75%</td>
</tr>
</tbody>
</table>

2.6.1 Methods

Materials

The second experiment was built using the same technical tools as Experiment 1 and used the same stimuli images and gesture videos. Participants were randomly assigned to one of the four conditions and a pseudo-randomised stimuli set containing two target meanings and associated distractors. The stimuli set consisted of two nouns, one from the set of ‘worn’ items (i.e. hat or scarf) and one from the set of ‘held’ items (i.e. cup and book). Each of these two nouns was then paired with one of the two dependents associated with the condition. For example, a stimulus set for a participant in one of the adjective conditions might

\textsuperscript{9}This experiment was granted ethical approval by the PPLS Research Ethics Committee. The study was also preregistered at https://shorturl.at/fGKR5 and all associated materials and code can be viewed at https://shorturl.at/jvHOZ.
consist of *striped hat* and *spotted cup*. The other three images in the 2x2 grid used in training and testing trials were chosen in the same way as in Experiment 1 to contrast both head noun and dependent.

**Procedure**

Participants were instructed that they were going to learn how to express ‘ownership of an item’ in a made-up sign language (genitive conditions) or that they would learn how to ‘describe an item’ (adjective conditions). Prior to the training phase, participants were exposed to the same kind of familiarisation trial as in Experiment 1. Following this, the main training phase took place and participants were told that a similar 2x2 grid would appear, but with one meaning highlighted, and that below the images they would see a video of person using gestures to convey the meaning in the highlighted image. All they had to do was sit back and watch carefully as each of the two target images were displayed with their corresponding gesture videos eight times each. Six of eight times the image would be described in the majority gesture order for that condition, and twice in the minority order. The training phase trials progressed automatically.

After this, participants were tested on what they had learned. They saw the same kind of image grid as in the training phase, but both possible gesture videos that corresponded to the target image were displayed under the images. These two videos looped simultaneously until participants chose one of them by clicking on it. Participants were instructed to ‘click on the corresponding gesture video’ like they had seen for those meanings during training. The testing phase had the same number of trials as the training phase (16) and participants saw both target meanings eight times and clicked a centred ‘Next’ button to proceed between trials. The location of the gesture videos (left or right) were randomised per trial per participant.

After the training and testing phases, participants were presented with translation trials, similar to Experiment 1 but twice, once for each target meaning. The gesture order they were prompted with for each target meaning was pseudorandomised so that one meaning appeared with a prenominal gesture order and one meaning with the postnominal one.\(^\text{10}\) Finally, participants answered the same demographics questions as in Experiment 1.\(^\text{11}\)

\(^{10}\)As in Experiment 1, details of translation trials and demographics questions can be found in the scripts in the OSF repository.

\(^{11}\)The most common reported L2s were the same as for Experiment 1. Full list of L2s reported by participants in Experiment 2 can be found in the OSF repository.
Participants

A total of 215 participants were recruited via the online crowdsourcing platform Prolific. We employed the same prescreening requirements as in Experiment 1. Participants were paid the equivalent of £9.50 per hour. We excluded 6 participants who stated that they were proficient in a sign language. A further 5 participants were excluded as they did not provide coherent responses to the translation and/or demographics questions (e.g. only included a random sequence of letters). Finally, 1 participant was excluded for pressing the same button over 90% of the time and data from 3 participants were lost due to technical issues during the experiment. After these preregistered exclusions there were 47 participants in the natural adjective condition, 50 in the unnatural adjective condition, 50 in the natural genitive condition, and 53 in the unnatural genitive condition.

2.6.2 Results

We had three main predictions for Experiment 2: (i) participants would show evidence of having learned the gesture orders they were trained on, by either reproducing and/or regularising the majority variant from their training. This prediction is a check to be sure that participants learn from the training data. Second (ii), we predicted that participants’ learning behaviour would be modulated by the naturalness of the majority variant in their condition: participants in the natural conditions were predicted to regularise more readily than participants in the unnatural conditions. Third (iii), we also predicted that participants would show an overall preference for natural orders by selecting more natural orders than predicted by chance across all conditions, regardless of the majority order. Finally, as in Experiment 1, we made the additional prediction that the naturalness preference would be stronger for adjectives than genitives (i.e., an interaction between naturalness and dependent type).

Learning

We first analyse whether participants generally learned the orders they were trained on, and whether this was modulated by condition as predicted. Figure 2.6 shows proportion choice of the majority orders in each condition (with right-hand panel collapsing across conditions) in the testing phase. We ran a mixed

\[12\] Note the difference in hourly payment was due to an increase in the minimum wage in Scotland between the two periods of data collection.
effects logistic regression model with *majority order* as the binary outcome variable (1 when participants’ choice matched the majority order they were trained on, and 0 when it did not), and fixed effects for *majority natural* (either natural or unnatural) and *dependent type* (either adjective or genitive) as well as their interaction. Both fixed effects were deviation-coded (genitive = 0.5 and adjective = -0.5, natural = 0.5 and unnatural = -0.5). The models also included a random slope for participants. The model had a significant positive intercept ($\beta = 1.51$, $SE = 0.13$, $z = 11.62$, $p < 0.001$) showing that, on average, participants across all conditions choose the majority order at a rate above chance. This confirms our first prediction, that participants generally learned from the gestures they were trained on. Model comparison using a likelihood ratio test revealed that the null model (reported above) was the best fit for the data, and that including *majority natural, dependent type* or their interaction did not improve model fit ($\chi^2 = 2.01$, $p = 0.16$; $\chi^2 = 2.91$, $p = 0.09$; $\chi^2 = 0.005$, $p = 0.94$). Indicating that there was no reliable difference in the likelihood of selecting majority orders for participants in the two natural conditions compared to the two unnatural condition, nor for participants exposed to adjectives or genitives.

![Figure 2.6](image_url)

Figure 2.6: Overall mean (right panel), conditions means (black circles, left panel), and individual participant proportions (coloured dots) of test trials where participants chose the majority input order for each condition. Error bars represent bootstrapped 95% CIs around the means. Dashed line shows chance level performance. Participants tended to produce more of the majority orders from their training than is expected by chance, and there was no difference between conditions.
Regularisation

The above analysis confirms that participants across all conditions were able to learn the order they were trained on, since they produced the majority order at a rate significantly above chance. However, this experiment was primarily designed to test participants’ regularisation behaviour across conditions. Here we quantify an increase in regularity in the system as a decrease in overall entropy between the input (training) and output (forced-choice selections) following, for example, Ferdinand et al. (2019), Motamedi et al. (2021), and Samara et al. (2017). The entropy ($H$) of a system is defined in this context as:

$$H(V) = - \sum_{v_i \in V} p(v_i) \log_2 p(v_i)$$

where, $(V)$ refers to the two possible gesture variants (prenominal and postnominal). All conditions had an input entropy of approximately 0.811, with a maximum possible entropy value of 1 (indicating that the output data shows an exactly 50/50 split between the two orders), and a minimum of 0 (indicating that the output data contains only a single order). The change in entropy was calculated by taking the output entropy of each participant, based on participants’ selections in the testing-phase, and subtracting the input entropy value for their condition. Figure 2.7 shows the mean entropy change in each condition (and collapsing across conditions). To evaluate if these changes are reliably greater than zero we calculated bootstrapped confidence intervals around the mean entropy changes for each condition. These were generated using the ‘boot’ package (Canty & Ripley, 2021) in R and based on 10,000 samples. These results were further supported by simulating 10,000 runs of the experiment with the probability of simulated participants choosing prenominal or postnominal order set to the input proportions during training (i.e. 0.75 for the majority order and 0.25 for the minority order). Z-scores were calculated based on the overall mean change in entropy between the observed experimental mean and the overall simulation mean. Similarly, individual z-scores were calculated for the mean change in entropy for each condition, compared to the corresponding simulation means. These analyses all indicate a reliable negative change between input and output entropy in each individual condition, and overall across conditions (see Table 2.5 for z-scores based on simulations). Importantly, CIs around differences in

---

13The necessarily non-linear nature of our entropy values made them unsuitable for analysis using linear models.
experimental means between conditions reveals no reliable differences between conditions in terms of regularisation behaviour (see Table 2.6).

Figure 2.7: Overall mean (right panel), conditions means (black circles, left panel), and individual participant values (coloured dots) for changes in entropy between input and output. Error bars represent bootstrapped 95% CIs around the means. The dotted line at 0 represents no change in entropy. Participant dots above this line signify an increase in entropy between input and output. There is an overall tendency towards greater regularity in participants’ outputs, and there was no difference between conditions in terms of regularisation behaviour.

Table 2.5: Comparison of experimental means and simulated means and resulting z-scores based on change in entropy across conditions for Experiment 2. Z-scores show that all experimental means are reliably different from the simulated means indicating that entropy dropped significantly in all conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Exp. mean</th>
<th>Sim. mean</th>
<th>z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>-0.221</td>
<td>-0.047</td>
<td>-13.13</td>
</tr>
<tr>
<td>Natural adjective</td>
<td>-0.169</td>
<td>-0.047</td>
<td>-4.59</td>
</tr>
<tr>
<td>Unnatural adjective</td>
<td>-0.167</td>
<td>-0.047</td>
<td>-4.53</td>
</tr>
<tr>
<td>Natural genitive</td>
<td>-0.243</td>
<td>-0.047</td>
<td>-7.44</td>
</tr>
<tr>
<td>Unnatural genitive</td>
<td>-0.299</td>
<td>-0.047</td>
<td>-9.51</td>
</tr>
</tbody>
</table>

Naturalness

Crucially, in addition to our predictions about learning and regularisation, we also predicted that if category-specific biases are active during learning, then natural orders (postnominal adjectives and prenominal genitives) would be chosen more than unnatural orders. Figure 2.8 shows the proportion of natural orders chosen
Table 2.6: Comparison of mean entropy change between conditions for Experiment 2. Includes 95% bootstrapped CIs around each mean. All intervals cross 0, showing no reliable differences between conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>$\overline{x}_a - \overline{x}_b$</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural adjective – Unnatural adjective</td>
<td>-0.003</td>
<td>-0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Natural adjective – Natural genitive</td>
<td>0.07</td>
<td>-0.09</td>
<td>0.24</td>
</tr>
<tr>
<td>Natural adjective – Unnatural genitive</td>
<td>0.13</td>
<td>-0.03</td>
<td>0.29</td>
</tr>
<tr>
<td>Unnatural adjective – Natural genitive</td>
<td>0.08</td>
<td>-0.08</td>
<td>0.23</td>
</tr>
<tr>
<td>Unnatural adjective – Unnatural genitive</td>
<td>0.13</td>
<td>-0.02</td>
<td>0.29</td>
</tr>
<tr>
<td>Natural genitive – Unnatural genitive</td>
<td>0.06</td>
<td>-0.10</td>
<td>0.22</td>
</tr>
</tbody>
</table>

by participants in each condition (right-hand panel collapsing across conditions). We ran a logistic mixed effects model on the binary outcome variable natural order (1 if the choice matched the predicted natural order, 0 otherwise). The rest of the model structure was identical to the one used to analyse learning behaviour above and thus included fixed effects for majority natural (either natural or unnatural) and dependent type (either adjective or genitive) as well as their interaction and a random slope for participants. The intercept term for the model was not significant ($\beta = -0.24$, $SE = 0.17$, $z = -1.36$, $p = 0.18$), indicating no overall preference for natural orders. A likelihood ratio test revealed that including the fixed effect of majority natural improved model fit compared to the null model ($\chi^2 = 115.07$, $p < 0.001$). There was a significant positive effect of majority naturalness ($\beta = 3.01$, $SE = 0.26$, $z = 11.64$, $p < 0.001$), showing that participants in the natural adjective and natural genitive conditions were more likely to select the natural order, compared to the grand mean. This is as expected since these were the orders participants were trained on. Including dependent type or the interaction between dependent type and majority natural did not improve model fit ($\chi^2 = 0.001$, $p = 0.97$; $\chi^2 = 6.12$, $p = 0.11$).

**Mutual information (exploratory)**

As is evident from Figure 2.7 there are a number of participants whose output in the testing phase shows a higher entropy than their training data. We ran an exploratory analysis (not included in our pre-registration) to see whether these participants might have had a different strategy for reducing unpredictable variation, which did not involve regularisation as we defined it above. For example, another way in which a system can become more consistent is by reducing the variants (orders) used for a particular meaning/lexical item (Samara et al., 2017;
Smith & Wonnacott, 2010; Smith et al., 2017). As participants were exposed to only two target meanings in Experiment 2, they might have conditioned the use of the two gesture orders on these two meanings. This strategy would result in an increase in overall entropy—as participants would use more variable order across the whole system—but a decrease in variability for a specific meaning. To capture this type of lexically-conditioned ordering, and disentangle it from overall entropy, we used a measure of Mutual Information of gesture order choice and meaning (lexical item). Mutual information is computed as:\(^{14}\)

\[ \text{Mutual information (MI)} = \text{overall entropy} - \text{conditional entropy (meaning)} \]

MI of 1 would indicate that participants perfectly condition the two gesture orders on the two meanings they are exposed to, whereas MI of 0 would indicate that participants do not make use of this strategy and that, instead, the

\[^{14}\text{Conditional entropy alone could not fully capture this behaviour since becoming more consistent across a whole system will also involve becoming more consistent within the context of a specific meaning, thus conflating the measure of regularisation and lexically-conditioned variation. Conditional entropy was defined as:} \]

\[ H(V|C) = - \sum_{c_j \in C} p(c_j) \sum_{v_i \in V} p(v_i|c_j) \log_2 p(v_i|c_j) \]
variability within each meaning reflects the variability of the system as a whole. The overall mean change in MI across all conditions is 0.12, although there is some variability between individual condition means (see Figure 2.9). Based on z-scores calculated between experimental and simulation means of change in MI, the increase in MI is consistent for all condition means except the natural adjective condition (see table 2.7). These results, in combination with the two entropy measures, show that some participants reduced unpredictable variation by using one gesture order more consistently across the whole system, whereas others maintained or even increased overall variability but made this variability predictable based on meaning.

Figure 2.9: Overall mean (right panel), condition means (black circles, left panel), and individual participant values (coloured dots, left panel) for changes in mutual information. Error bars represent bootstrapped 95% CIs around the means. The dashed line at 0 represents no change in MI between training and output. Space above the dashed line represent space of possible MI values. All measures are reliably different from 0, except the change in MI in the natural adjective condition.

Table 2.7: Experimental means, simulated means and resulting Z-scores for change in mutual information

<table>
<thead>
<tr>
<th>Condition</th>
<th>Exp. mean</th>
<th>Sim. Mean</th>
<th>z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.120</td>
<td>0.055</td>
<td>12.27</td>
</tr>
<tr>
<td>Natural adjective</td>
<td>0.064</td>
<td>0.055</td>
<td>0.74</td>
</tr>
<tr>
<td>Unnatural adjective</td>
<td>0.120</td>
<td>0.055</td>
<td>6.18</td>
</tr>
<tr>
<td>Natural genitive</td>
<td>0.186</td>
<td>0.055</td>
<td>12.41</td>
</tr>
<tr>
<td>Unnatural genitive</td>
<td>0.109</td>
<td>0.055</td>
<td>5.03</td>
</tr>
</tbody>
</table>
2.6.3 Experiment 2: Discussion

In Experiment 2 participants learned a variable system of word order for adjective or genitive meanings where we manipulated whether the input data they were trained on mainly consisted of natural or unnatural orders. We measured how well participants learned the system they were trained on, the extent to which they regularised the use of one gesture order, and more importantly, their use of natural orders. The results showed that participants learned and regularised the gesture systems they were trained on. However, contrary to our prediction, participants’ learning and regularisation behaviour was not modulated by the naturalness of the majority variant that they were trained on. Participants in the natural conditions did not learn the language systems more accurately, nor did they regularise the majority variant more readily than participants in the unnatural conditions. Similarly, we did not find evidence of any general preference for natural orders across conditions: participants were overall just as likely to choose natural and unnatural orders. Finally, as in Experiment 1, we found no evidence that a preference for naturalness targeted adjectives more than genitives: the interaction between naturalness and dependent type was not significant.

These results thus do not support the hypothesis that category-specific biases affect linguistic behaviour during the learning of an ordering system, at least in the case of these two dependent-types. However, the reliable reduction in entropy across all conditions, combined with the reliable increase in Mutual Information based on head noun in three of the four experimental conditions, adds to the body of literature showing that learners are biased against unpredictable variation (Hudson Kam & Newport, 2009; Samara et al., 2017; Smith & Wonnacott, 2010; Smith et al., 2017). Rather than reproducing the unconditioned variation from training, participants tended to over-extend the use of one order at test (regularisation) or condition the use of the two orders on some lexical aspect of the input.

2.6.4 General Discussion

Previous experimental work has found evidence suggesting that cognitive biases, which are active at the level of the individual, may help to explain typological patterns (e.g. Culbertson et al., 2012; Finley, 2018; Martin et al., 2020). For example, there is a typological trend for harmonic word order in the noun phrase (e.g., consistent ordering of dependents before or after nouns), and participants in artificial language learning experiments prefer exactly these orders. In this study we targeted an exception to the typological trend towards harmonic word
order in the noun phrase: a non-harmonic order of adjectives and genitives, with adjectives after the noun and genitives before, is just as common as the two harmonic patterns. We sought to explore whether this pattern might be caused by conflicting category-specific biases influencing individuals’ ordering preferences for adjectives and genitives respectively. Previous experimental work has mainly found that category-specific biases tend to be active in contexts where participants have little or limited evidence for a wider language system (Culbertson, Schouwstra, & Kirby, 2020; Martin et al., 2020; Schouwstra & de Swart, 2014) whereas more system-wide biases—like harmony—are active when these language systems are in place and participants are tasked with learning them (Culbertson et al., 2012; Samara et al., 2017; Smith & Wonnacott, 2010). The experiments in this study investigate both of these contexts. If category-specific biases are also active during learning, this would provide more opportunity for these biases to influence typological structure and, potentially, compete with the system-wide bias for harmony.

The experiments reported here were thus designed to investigate two main questions. First, whether category-specific word order preferences for adjectives and genitives influence ordering preferences in the absence of a linguistic system. Second, whether these preferences in turn influence learning and regularisation of a (miniature) gestural linguistic system. Experiment 1 showed that participants have clear preferences when asked to select a gesture order expressing either a descriptive (adjective) or a possessive (genitive) meaning without any wider linguistic structure or system being provided to them. Participants in the adjective condition tended to select postnominal orders, whereas participants in the genitive condition tended to select prenominal orders. These preferences closely follow the ordering preferences seen in typological data of both spoken and signed languages (Coons, 2022; Dryer, 2013a). The main difference between our results and the patterns seen in (spoken) language typology is that the postnominal adjective preference was not stronger than the prenominal genitive preference. These results are in line with previous research which has found evidence for a postnominal adjective preference based in experimental tasks (Culbertson, Schouwstra, & Kirby, 2020; Jaffan et al., 2020), but expands on this by providing the first experimental evidence for a prenominal genitive preference.

By contrast, Experiment 2 found no evidence for these category-specific biases when participants were tasked with learning an ordering system to express the meaning of one of these dependent types. Participants trained on a more natural system, where the majority order aligned with the preferred orders in Experiment 1, did not reproduce the majority orders more faithfully, or regularise these orders...
more readily than if they were trained on an unnatural system. Overall, learning and regularisation behaviour was comparable across all four conditions. Thus, there was no tendency for participants to produce more natural orders than what would be expected by chance: the overall input proportion of natural orders across all conditions was 50%, and this remained the case in participants’ output across conditions. In other words, the natural order was only used more than the unnatural order if it was the majority one for a given participant.

To summarise, we have evidence that category-specific ordering preferences for these two dependent types influence linguistic behaviour in a specific context. In particular, when participants must choose the order of an individual linguistic item, in the absence of any knowledge of the wider system that that item belongs to. Once items are taught to participants as part of a system, we no longer found evidence for these category-specific preferences. Instead, systems that align with or deviate from natural orders were learned equally well.

**Revealing category-specific biases**

These results differ from previous work on which found that preferences for basic word order conditioned on event type (i.e. SOV for extensional vs. SVO for intensional) influences behaviour both in the absence of a system, and during learning of a system (Motamedi et al., 2022). This difference is especially interesting in light of the potential conceptual parallel between the preferences found here and those found for event type. In intensional events, the existence of the object depends on the action of the verb (e.g. *gnome dreams of banana*). Similarly, some adjectives (such as scalar adjectives) depend on the head noun for their interpretation. In both cases, there is a preference for ordering the dependent element (the syntactic object, or adjective) after the element on which is depends (the verb or noun). Conversely, objects of extensional events and possessors are both more independent of their heads; the object of an extensional verb exists independent of it, and a possessor does not typically rely on the head noun for its interpretation. There is nothing about the concept of *cup* which changes the way we interpret the existence of *vampire in vampire’s cup*. Given this parallel, it is striking that there is a difference between our results, which demonstrate a lack of naturalness preference in learning as opposed to a one-off judgement, compared to those reported in (Motamedi et al., 2022), which do show this preference in both tasks.

One possibility is that the biases which affect adjectives and genitives are relatively weaker than those which govern the event type conditioning. Assum-
ing that a category-specific bias must be relatively strong in order to overcome evidence of a conventionalised order, this could explain why only the latter are evident in learning. However, this would lead us to expect more languages that condition basic word order on event type compared to languages which have non-harmonic orders for adjectives and genitives. This is clearly not the case. In fact, conditioning word order on event type seems to be comparatively rare in typology, although more detailed research is needed (Flaherty et al., 2018; Napoli et al., 2017). However, it could be that this pattern does not survive because the pressure to regularise and converge on a consistent basic word order is very strong (and stronger than the pressure to converge on a single noun phrase order). Many languages use fixed word order to signify who does what to whom. In such languages a conventionalised basic word order is crucial for communicative purposes. This pressure could out-compete the category-specific preferences to condition order on event type, making it rare typologically even if it is evident in experimental contexts.

Of course, this brings up the question of how to explain the typological prevalence of the non-harmonic order aligned with the category-specific biases we have found here. If these biases are present only during language emergence, but not during learning—either because they are too weak, or because the mechanism which underlies them is simply not active during learning—then it is somewhat surprising that these biases have persisted in the face of a competing pressure for harmony.

In light of this, it is worth considering the possibility that contexts similar to language emergence might be present, albeit to a lesser degree, in more typical linguistic contexts. In other words, we might imagine that category-specific biases more generally, or at least for these particular dependent types, arise whenever some level of linguistic innovation or creativity is required. Although adults tend to be fully proficient language users, there is still a surprising level of novelty and innovation employed in everyday language tasks (Christiansen & Chater, 2022). This includes instances where we integrate new lexical items into pre-existing categories and, more specifically, during language acquisition children often have to produce structures for which they have no direct evidence (Chomsky, 1972; Perfors et al., 2011). Support for the possibility that tasks which involve some innovation might be more likely to reveal category-specific biases can be seen in, for example, studies examining biases for noun phrase homomorphism and affix ordering (Culbertson & Adger, 2014; Martin et al., 2019, 2020; Saldana et al., 2021). In these experiments, ordering biases emerge when learners have been trained on part of a system but must extrapolate beyond their input in
the critical task. For example, participants in Martin et al. (2020) learn that a single dependent comes after the noun, but must extrapolate beyond that to generate the relative order of multiple dependents. It is possible that the biases observed for adjectives and genitives in Experiment 1 would also influence linguistic behaviour under such conditions. Lack of direct evidence for the full language system might cause participants to “fall back” on these category-specific biases to a certain degree. We hope to examine this possibility in future studies. If participants show a tendency towards natural orders in both extrapolation and improvisation/emergence contexts, then category-specific biases would have more opportunity to impact linguistic behaviour and therefore shape typology. This would be in line with theories stipulating that innovation and creativity have continuous effects on language, as they are common mechanisms in everyday language use (Christiansen & Chater, 2022).

**Why these category-specific biases?**

The experiments presented here do not directly test the underlying cause for the category-specific preferences we have identified. Nevertheless, above we mentioned a possible parallel between basic word order and noun phrase word order. Specifically, the idea that adjectives, like the objects in intensional events, depend on the head for their interpretation. The possibility that adjectives might tend to be postnominal for this reason is explicitly discussed by Culbertson et al. (2012), and supported by the results of several experiments (Culbertson, Schouwstra, & Kirby, 2020; Jaffan et al., 2020; Rubio-Fernandez et al., 2022; Weisleder & Fernald, 2009). Specifically, many common adjectives depend on the context of the noun in order for their meaning to be correctly interpreted. For example, comparing the meaning of the adjective *good* in the phrases *good pianist* and *good food* shows that the adjective denotes two very different properties of the head noun in each case. In the first phrase it concerns how well the musician plays their instrument, whereas in the second phrase it refers to some property of the food being considered tasty. For languages with postnominal adjectives, like Thai and Navajo, the noun has already been encountered when the adjective must be interpreted, allowing for incremental semantic processing. By contrast, users of languages that have prenominal adjectives cannot interpret the meaning of these adjective as soon as they are encountered, but need to keep it them in memory and interpret them once the head noun has given the relative context.

Evidence from typology and language processing also provide some potential explanations for grounding of the prenominal preference for genitives. In partic-
ular, this preference might be rooted in the association between ownership and animacy. Prototypical possessors tend to be high on the animacy scale (Rosenbach, 2008; Silverstein, 1986; Yamamoto, 1999) and animate entities have been argued to hold a privileged position in language processing by virtue of being highly accessible (Dahl, 2008). This may lead to such referents being linearised earlier in a linguistic construction (R. Hawkins, 1981; McDonald et al., 1993; Tanaka et al., 2011). The privileged position of animate entities has been suggested as an explanation for why there is a cross-linguistic prevalence of subject initial languages in both spoken and signed languages (Dryer, 2013c; Napoli & Sutton-Spence, 2014). This preference has also been found in silent gesture studies (Goldin-Meadow et al., 2008; Schouwstra & de Swart, 2014), and in some young sign languages (Meir et al., 2017; Sandler et al., 2005). With respect to genitives, several languages with variable genitive order, like English, Dutch and Low Saxon, condition their use of prenominal versus postnominal genitive order on the animacy of the possessor, such that prenominal order is used for animate possessors and postnominal order is used for inanimate possessors (Rosenbach, 2005; Strunk, 2004; van Bergen, 2011).

We cannot directly provide evidence for any of these explanations here. However, more targeted studies examining the proposed cause of such category-specific biases would be a fruitful way to expand our understanding of the cognitive grounding of these preferences.

## 2.7 Conclusion

This study explored the role of category-specific and system-wide biases on language structure. Category-specific biases target individual words, phrases, or utterances; system-wide biases describe features that hold across these. Here we explore a case where these two types of biases appear to be in conflict: a system-wide preference for harmony in the noun phrase (i.e., consistent order of nouns and dependents) and category-specific biases that favour pre-nominal placement of genitives but post-nominal placement of adjectives. We were interested in the contexts under which such category-specific biases might influence language, and thus push against system-wide preferences like harmony. We found evidence for both these category-specific biases when participants were asked to judge word order in gesture sequences in the absence of a wider linguistic system. However, these biases did not modulate learning of a linguistic system. These results suggest that category-specific biases may play a role in shaping language most
clearly in contexts that require innovation of expressions of meanings, rather than acquisition of conventionalised expressions of meanings. System-wide biases in favour of harmony (and regularity more generally) may instead be the strongest guiding forces during learning tasks. These results leave open whether there are additional contexts in which both pressures are at play. For example tasks which involve substantial extrapolation beyond the learned input may be the locus of direct competition between the category-specific and system-wide biases.
References


Appendix

2.A Outline

This appendix includes further information about additional data collected in Experiments 1 and 2 as well as details regarding various design choices that apply throughout the thesis. Section 2.B contains detailed explanations for the motivation behind the stimuli selection and stimuli design, as well as the full set of meanings/images used across experiments in this thesis. Section 2.C gives details about the consent information and instructions that participants were shown in both of the experiments. Section 2.D contains tables summarising the answers that participants gave to the questionnaires at the end of the experiments. Finally, 2.G describes the set-up and results of the Monte Carlo simulation that was used to evaluate the reliability of the changes in entropy and Mutual Information that were observed in Experiment 2 in Chapter 2.

2.B Stimuli design

The following section outlines the creation of the two sets of visual stimuli employed in both of the experiments and the motivation for why the specific meanings were used as stimuli.

2.B.1 Image stimuli creation

The nouns represented in the image stimuli consisted of two sets of items, the “worn” items were hat and scarf and the “held” items were cup and book. These acted as the head nouns in the intended meanings participants would encounter in the stimuli images. These nouns were combined with either of the two descriptive (adjective) elements, namely spotted and striped, and/or with either of the two possessor (genitive) elements, vampire and cyclops. In total this resulted in 16 different images which can be seen in Figures 2.B.1 and 2.B.2). These images were
drawn by hand as vector images using Inkscape, and exported to png format for inclusion in the experiments.
Figure 2.B.2: All image stimuli for adjective meanings.
Video stimuli creation

Video stimuli consisted of 32 unique videos, each showing a model gesturer producing a sequence of two gestures. The videos were constructed in pairs such that two gesture sequences denoted the same meaning, one where the first gesture denoted the head noun and one where the first gesture denoted the dependent (adjective or genitive element). The gestures were produced to the sound of a metronome to ensure that gestures started and ended at the same time. All videos started and ended with the gesturer in a neutral position, with both hands clasped loosely in front of them, and they maintained a neutral facial expression throughout. Several sequences of the same two gestures were produced in a row, and pairs of gestures (prenominal and postnominal) were matched for simultaneity in movement by playing the videos side-by-side in iMovie. None of the videos included other movement artefacts, such as blinking. The videos were edited to be exactly 4,389 milliseconds long using HandBreak, allowing them to loop and remain synchronised. Sample stills from two pairs of gesture videos can be seen in Figure 2.B.3.

Figure 2.B.3: Sample stills from gesture videos showing sequences of two gestures for both adjective (a-b) and genitive (c-d) meanings.
2.B.3 Stimuli motivation

Several important factors were considered when designing this stimuli to ensure they represented optimal visual imagery for the study, as research has shown that subtle aspects of the stimuli can affect ordering biases in silent gesture studies (Kirton et al., 2021). Firstly, the items (hat, scarf, cup and book) were chosen as they represent distinct physical items which could be easily and naturally denoted with two-handed gestures in a centralised position on the body. This was done to avoid potential ordering effects due to some gestures being one-handed and others being two-handed. Similarly, the centralised location of each of the items on the body avoided potential left-right biases between gestures denoting those items. There was also a potential for a “top-down” bias, especially for any combination of meanings including the hat meaning, as the gesture denoting this meaning was localised on the head. To examine these potential motivations there was a question with a text-based response in Experiment 1 where participants could specify the motivations for their order choices. Details about this question can be found in section 2.D.

The genitive meanings, vampire and cyclops, were chosen as possessors needed to be denoted through gestures which signified inalienable aspects of the characters. Gestures denoting individuals often pick out a specific physical characteristic of that person. For example, using a gesture referring to an eye-patch to denote the meaning/individual pirate (Schouwstra & de Swart, 2014). Crucially, I wanted to use characters where the default denoting gestures did not refer to things that the possessors owned or wore. Referring to an individual by items they own would have created a type of “double-possessive” meaning within the gesture phrases when these individuals were part of a possessive relation with another noun. Thus I chose vampire and cyclops since their defining characteristics (the sharp teeth and singular eye respectively) are inalienable parts of these characters. This creates a distinction in the way the gestures denote an individual and the way they denote the items these individuals possess.

Finally, the specific adjectives were chosen as they had been successfully employed in improvisation-based silent gesture research before (see e.g. Culbertson et al., 2020). Pattern adjectives are comparatively easy to depict in a clear and iconic way, compared to more abstract descriptive features like colour or surface texture.
2.C Consent and instructions

Experiment 1 included a web-page where participants could indicate their consent to take part in the study, as well as a link to an external (pdf) information sheet with specific information about data storage, right to withdraw etc. Below is the text included on the consent page for both the adjective and genitive conditions (differences between conditions are bolded). Copies of the full information sheet provided to participants can be found on the OSF pages linked in the main content chapter.

2.C.1 Experiment 1: Consent info

Adjective condition

Welcome!

This is a study about how you describe items in a sign language. It will take between 3-5 minutes to complete and you will be paid £0.75 for your time.

The study is being conducted by Drs Jennifer Culbertson and Simon Kirby at the University of Edinburgh and Dr. Alexander Martin at the University of Paris. The study has been granted ethical approval. Please click here to read an information letter (pdf) about the study.

By clicking on the agree button you indicate that:
- You are a native speaker of English, at least 18 years old.
- You have read the information letter.
- You voluntarily agree to participate, and understand you can stop your participation at any point.
- You agree that your anonymous data may be kept permanently in Edinburgh University Archives and may be used for research purposes.

If you do not agree to all of the above, please exit the experiment now.

The experiment plays VIDEO and we have checked your browser to make sure it will play. You are all set!

Genitive condition
Welcome!

This is a study about how you express ownership in a sign language. It will take between 3-5 minutes to complete and you will be paid £0.75 for your time.

The study is being conducted by Drs Jennifer Culbertson and Simon Kirby at the University of Edinburgh and Dr. Alexander Martin at the University of Paris. The study has been granted ethical approval. Please click here to read an information letter (pdf) about the study.

By clicking on the agree button you indicate that:
- You are a native speaker of English, at least 18 years old.
- You have read the information letter.
- You voluntarily agree to participate, and understand you can stop your participation at any point.
- You agree that your anonymous data may be kept permanently in Edinburgh University Archives and may be used for research purposes.

If you do not agree to all of the above, please exit the experiment now.

The experiment plays VIDEO and we have checked your browser to make sure it will play. You are all set!

2.C.2 Experiment 1: Instructions

Adjective condition
Contrast trial instructions: The two set of instructions below vary in the order of the words “items” and “patterns”. This was done to examine if instruction order correlated with order choice. No evidence was found in support of this.

1. On the next screen you will see four images, each representing a different instance of ownership. Across the images both the patterns and the items vary. Study the images, then continue to the next screen. Click “Next” to continue.

2. On the next screen you will see four images, each representing a different instance of ownership. Across the images both the items and
the patterns vary. Study the images, then continue to the next screen. Click “Next” to continue.

Forced-choice instructions: These instructions preceded the trial where participants had to choose a gesture order to express a target meaning.

You will now see the same four images again but one will be highlighted. Next to these images there will be two videos. These show two different ways in which the item in the highlighted image could be described in a made-up sign language. Your task is to choose the video you think would best describe the item. Study both videos, then make your selection. Click “Next” to begin.

Slider instructions: These instructions preceded the slider trial where participants had to indicate how strong their preference was for the video they had chosen.

Please indicate how strong your preference is for the video you chose in the previous trial. Use the slider to indicate the strength of your preference by dragging it towards the video you prefer.

Translation instructions: These instructions preceded the translation trial where participants had to provide an English translation for the meaning displayed in their chosen gesture video.

Here is the sign language option you chose. How would you translate this expression into English?

Motivation prompt: The question used to prompt participants free-form response for why they preferred a given video.

Why did you prefer this video?

Questionnaire prompts: The questions given to participants at the end to the experiment.

1. Do you have any knowledge of a sign language? (If yes, please state which).

2. What spoken languages do you know and to what proficiency? Please indicate proficiency on a scale of 1-10 (e.g. English - 10, Tamil - 8, French - 4).

Genitive condition
Contrast trial instructions: The two set of instructions below vary in the order of the words “items” and “owners”. This was done to examine if instruction order correlated with order choice. No evidence was found in support of this.
1. *On the next screen you will see four images, each representing a different instance of ownership. Across the images both the owners and the items vary. Study the images, then continue to the next screen. Click “Next” to continue.*

2. *On the next screen you will see four images, each representing a different instance of ownership. Across the images both the items and the owners vary. Study the images, then continue to the next screen. Click “Next” to continue.*

Forced-choice instructions: These instructions preceded the trial where participants had to choose a gesture order to express a target meaning.

*You will now see the same four images again but one will be highlighted. Next to these images there will be two videos. These show two different ways in which ownership of the item in the highlighted image could be expressed in a made-up sign language. Your task is to choose the video you think would best convey this meaning. Study both videos, then make your selection. Click “Next” to begin.*

Slider instructions: Identical to the adjective condition.
Translation instructions: Identical to the adjective condition.
Motivation prompt: Identical to the adjective condition.
Questionnaire prompts: Identical to the adjective condition.

**2.C.3 Experiment 2: Consent info**

Experiment 2 included a web-page where participants could indicate their consent to take part in the study. This page also displayed the information sheet with specific information about data storage, right to withdraw etc. Copies of the full information sheet provided to participants can be found on the OSF pages linked in the main content chapter.

**2.C.4 Experiment 2: Instructions**

Below is a transcript of all the instructions that participants were given during Experiment 2 (differences between adjective conditions and genitive conditions are bolded for clarity).

**Adjective conditions**
Contrast trial instructions: Instructions provided at the start of the natural adjective and unnatural adjective condition.
In the first part of the experiment, you will observe gestures being produced which correspond to the meaning of different images. The gestures show how you describe an item in a made-up sign language. On the next screen you will see an example of the kinds of images you will see in the rest of the experiment. The images vary by the item they show and the pattern. Study the images and then continue to the next screen.

Training instructions: Instructions provided before the training phase across all conditions.

In the first part of the experiment, one of the images in the grid will be highlighted and underneath the images you will see a video of a person using gestures to convey the meaning of the highlighted image. These kinds of images, along with their accompanying gestures, will be shown to you several times. Just sit back and watch carefully.

Testing instructions: Instructions provided before the testing phase across all conditions.

In this second part of the experiment, you will be shown the same collection of images again as well as two gesture videos. Please click on the corresponding gesture video like you saw in the first part. This will repeat several times for each image.

Translation instructions: Instructions provided before the translation trials across all conditions.

You will now be asked to translate some of these gesture sequences into English, please type your answer in the box for each question.

Questionnaire prompts: Instructions and prompts presented during the questionnaire across all conditions.

Please answer the following questions:

1. Do you have any knowledge of a sign language? (If yes, please state which).

2. What spoken languages do you know and to what proficiency? Please indicate proficiency on a scale of 1-10 (e.g. English - 10, Tamil - 8, French - 4).
Genitive conditions
Contrast trial instructions: Instructions provided at the start of the natural genitive and unnatural genitive condition.

In the first part of the experiment, you will observe gestures being produced which correspond to the meaning of different images. The gestures show how you express ownership of an item in a made-up sign language. On the next screen you will see an example of the kinds of images you will see in the rest of the experiment. The images vary by the item they show and the owner. Study the images and then continue to the next screen.

Training instructions: Identical to the adjective conditions.
Testing instructions: Identical to the adjective conditions.
Translation instructions: Identical to the adjective conditions.
Questionnaire prompts: Identical to the adjective conditions.

2.D Language data
Several aspects of the questionnaire data from Experiment 1 and 2 were hand-coded and/or summarised. First, Table 2.D.1 includes all the additional spoken languages that participants in Experiment 1 reported having knowledge of. The same data for Experiment 2 is found in Table 2.D.2. The most common additional spoken languages across both experiments were French, German, and Spanish, probably a reflection of these being common modern languages taught at school in many English-speaking countries. Counts of languages exceed the total number of participants for both experiments since many participants knew several languages. Languages are sorted from most frequent to least frequent, with alphabetical ordering for languages with the same number of speakers. The additional column in Table 2.D.1 displays which of these languages are reported as having majority postnominal adjective order based on data from WALS. Languages with value NA either had no classification in WALS or were classified as having no dominant order. This data was used in supplementary analyses of the adjective condition results in Chapter 2 Experiment 1 to examine if knowledge of a postnominal adjective language affected the probability of postnominal adjective selection in the main forced-choice task.
Table 2.D.1: Languages (other than English) spoken by participants whose data was included in Experiment 1 in Chapter 2.

<table>
<thead>
<tr>
<th>Languages</th>
<th>N participants</th>
<th>N-Adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>79</td>
<td>Y</td>
</tr>
<tr>
<td>Spanish</td>
<td>55</td>
<td>Y</td>
</tr>
<tr>
<td>German</td>
<td>34</td>
<td>N</td>
</tr>
<tr>
<td>Afrikaans</td>
<td>13</td>
<td>NA</td>
</tr>
<tr>
<td>Italian</td>
<td>8</td>
<td>Y</td>
</tr>
<tr>
<td>Zulu</td>
<td>8</td>
<td>Y</td>
</tr>
<tr>
<td>Irish</td>
<td>7</td>
<td>Y</td>
</tr>
<tr>
<td>Japanese</td>
<td>6</td>
<td>N</td>
</tr>
<tr>
<td>Arabic</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>Dutch</td>
<td>5</td>
<td>N</td>
</tr>
<tr>
<td>Welsh</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>Portuguese</td>
<td>4</td>
<td>Y</td>
</tr>
<tr>
<td>Punjabi</td>
<td>4</td>
<td>NA</td>
</tr>
<tr>
<td>Other*</td>
<td>5</td>
<td>NA</td>
</tr>
<tr>
<td>Mandarin</td>
<td>3</td>
<td>N</td>
</tr>
<tr>
<td>Russian</td>
<td>3</td>
<td>N</td>
</tr>
<tr>
<td>Urdu</td>
<td>3</td>
<td>N</td>
</tr>
<tr>
<td>Xhosa</td>
<td>3</td>
<td>Y</td>
</tr>
<tr>
<td>Cantonese</td>
<td>2</td>
<td>N</td>
</tr>
<tr>
<td>Greek</td>
<td>2</td>
<td>N</td>
</tr>
<tr>
<td>Hindi</td>
<td>2</td>
<td>N</td>
</tr>
<tr>
<td>Polish</td>
<td>2</td>
<td>N</td>
</tr>
<tr>
<td>Shona</td>
<td>2</td>
<td>Y</td>
</tr>
<tr>
<td>Sotho</td>
<td>2</td>
<td>Y</td>
</tr>
<tr>
<td>Catalan</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>Filipino</td>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>Gaelic</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>Gujarati</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>Hebrew</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>Indonesian</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>Konkani</td>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>Korean</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>Kriolu</td>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>Languages</td>
<td>N participants</td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>Latin</td>
<td>1 Y</td>
<td></td>
</tr>
<tr>
<td>Mandinka</td>
<td>1 Y</td>
<td></td>
</tr>
<tr>
<td>Nyanja</td>
<td>1 Y</td>
<td></td>
</tr>
<tr>
<td>Romanian</td>
<td>1 Y</td>
<td></td>
</tr>
<tr>
<td>Swedish</td>
<td>1 N</td>
<td></td>
</tr>
<tr>
<td>Tagalog</td>
<td>1 NA</td>
<td></td>
</tr>
<tr>
<td>Wolof</td>
<td>1 Y</td>
<td></td>
</tr>
</tbody>
</table>

*includes counts of unclear language labels

Table 2.D.2: Languages (other than English) spoken by participants whose data was included in Experiment 2 in Chapter 2.
Participants in both experiments produced translations for the meanings they were exposed to, prompted by gesture videos. In Experiment 1 there was only one such translation trial, whereas in Experiment 2 there were two of them. The responses to these translations were hand-coded for order of elements in the English translation. The coding scheme had three levels *prenominal*, *postnominal*, or *NA*. Table 2.E.1 shows the number of responses that fall into these categories for translations of adjective and genitive meanings based on data from Experiment 1. The same data for Experiment 2 can be seen in table 2.E.2. Note that these results were used in an exploratory analysis in Experiment 1, examining if the orders in the English translations matched the ordering preferences based on video choice (i.e. did participants in the genitive condition produce more prenominal orders than is expected by chance, and did participants in the adjective condition produce more postnominal orders than is expected by chance). The results of these analyses is that this is indeed the case for the genitive condition ($\beta = 1.17$, $SE = 0.28$, $z = 4.15$, $p < 0.001$) but not the adjective condition ($\beta = 0.27$, $SE = 0.24$, $z = 1.13$, $p = 0.26$). Overall, more responses could be classified as either prenominal or postnominal when participants were translating adjective
meanings than when they were translating genitive meanings.

Table 2.E.1: Order of English translations given for target meanings in Experiment 1.

<table>
<thead>
<tr>
<th>Order</th>
<th>Adjective</th>
<th>Genitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postnominal</td>
<td>71</td>
<td>30</td>
</tr>
<tr>
<td>Prenominal</td>
<td>54</td>
<td>87</td>
</tr>
<tr>
<td>NA</td>
<td>35</td>
<td>43</td>
</tr>
<tr>
<td>Total</td>
<td>160</td>
<td>160</td>
</tr>
</tbody>
</table>

Table 2.E.2: Order of English translations given for target meanings in Experiment 2 in Chapter 2.

<table>
<thead>
<tr>
<th>Order</th>
<th>Adjective</th>
<th>Genitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postnominal</td>
<td>48</td>
<td>15</td>
</tr>
<tr>
<td>Prenominal</td>
<td>111</td>
<td>77</td>
</tr>
<tr>
<td>NA</td>
<td>35</td>
<td>114</td>
</tr>
<tr>
<td>Total</td>
<td>194</td>
<td>206</td>
</tr>
</tbody>
</table>

In addition to this categorisation of translation order, the type of phrase used to denote the meanings in the translation trials was also coded. The coding scheme for this differed between translations of adjective and genitive meanings. For genitive meanings the categories were *verb phrase* (these include possessive verb phrases like “the cyclops has a book” and action-based phrases like “the vampire wears a hat”), *genitive* (including both the *s* possessive and *of* possessive), *prepositional phrase* (including “cyclops with a cup” and “a hat on a vampire”) and *other* (including all phrases which could not be categories as any of the previous categories, such as the use of plain juxtaposition “hat vampire”). For adjective meanings the categories were *adjective* (including phrases like “spotted hat” and “stripy book”), *prepositional phrase* (like “hat with spots” or ”stripes on a book”), *verb phrase* (including “the cup has spots”), and *other* (including all phrases which could not be categories as any of the previous categories, such as the use of plain juxtaposition “scarf stripes”). Summary data of the number of translations that fell into each category for adjective and genitive meanings in Experiment 1 can be seen in Table 2.E.3, and the same data is available for Experiment 2 in Table 2.E.4. The translations that participants gave varied quite widely in both experiments, but more adjective meanings were straightforwardly
translated using English adjective phrases than genitive meanings were translated into s possessives or of possessives. While the comparatively low number of pure genitive translations that participants provided may seem to undermine the idea that participants activated a possessive representation for those stimuli items, there were many instances of possessive verb phrases within the high number of verb phrase responses to this translation question. Additionally, while these translations acted as an attempt to monitor the meanings that were activated in participants’ minds while participating in this task, the act of translating these meanings into English adds an additional step of transformation to those representations. Therefore we cannot take these translations as true reflections of participants’ representations of these meanings during the experimental tasks.

Table 2.E.3: Classification of English translations given for target meanings in Experiment 1 Chapter 2

<table>
<thead>
<tr>
<th>Class</th>
<th>Adjective</th>
<th>Genitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>54</td>
<td>NA</td>
</tr>
<tr>
<td>Genitive</td>
<td>NA</td>
<td>10</td>
</tr>
<tr>
<td>Preposition</td>
<td>44</td>
<td>21</td>
</tr>
<tr>
<td>Verb phrase</td>
<td>1</td>
<td>54</td>
</tr>
<tr>
<td>Other</td>
<td>61</td>
<td>75</td>
</tr>
<tr>
<td>Total</td>
<td>160</td>
<td>160</td>
</tr>
</tbody>
</table>

Table 2.E.4: Classification of English translations given for target meanings in Experiment 2 Chapter 2

<table>
<thead>
<tr>
<th>Class</th>
<th>Adjective</th>
<th>Genitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>106</td>
<td>NA</td>
</tr>
<tr>
<td>Genitive</td>
<td>NA</td>
<td>4</td>
</tr>
<tr>
<td>Preposition</td>
<td>37</td>
<td>41</td>
</tr>
<tr>
<td>Verb phrase</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>Other</td>
<td>48</td>
<td>145</td>
</tr>
<tr>
<td>Total</td>
<td>194</td>
<td>206</td>
</tr>
</tbody>
</table>

2.F Motivation data

Finally, Experiment 1 also included a question allowing participants to express the motivation for why they chose the gesture order they did. The data was clas-
sified into four categories, *language motivated* was used when the reply contained specific reference to orders in participants’ native language(s). *Intuition* was used for all instances where participants expressed some notion of an unmotivated preference (e.g. “it felt more clear”). *Directional* was used for all motivations that indicated that the location of the gestures themselves motivated the video choice (e.g. a “top down” preference). Finally, the last category was *Other*, and was used for all other motivations which did not fall into the previous three categories. Summaries of this data from Experiment 1 can be seen in Table 2.F.1. The most common motivation was intuition for both conditions. There were more language motivated selections in the adjective condition, and more directional motivations in the genitive condition. This might be related to the ease with which participants interpreted the meaning of the images, such that participants found it easier to extract an adjective meaning from the adjective stimuli, than a genitive meaning from the genitive stimuli. With a more direct parallel in place between the gestures and participants’ native language grammar in the adjective condition this could explain the presence of more language motivated choices.

Table 2.F.1: Motivations given for gesture order selection provided by participants in the adjective and genitive conditions in Experiment 1 in Chapter 2.

<table>
<thead>
<tr>
<th>Motivation type</th>
<th>N participants (adjective)</th>
<th>N participants (genitive)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intuition</td>
<td>159</td>
<td>161</td>
</tr>
<tr>
<td>Language</td>
<td>25</td>
<td>7</td>
</tr>
<tr>
<td>Directional</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Other</td>
<td>6</td>
<td>15</td>
</tr>
</tbody>
</table>

2.G Monte Carlo simulation

In this section I will give some further details about the set-up of the Monte Carlo simulation that I used to validate the reliability of the changes in overall entropy and Mutual Information (MI) that were calculated based on the data in Experiment 2 (see Chapter 2 section 2.6.2). This simulation was used to compare the observed experimental values for change in overall entropy and change in MI based on head noun, to a theoretical baseline of sampling based on input probability.

Monte Carlo simulations employ repeated sampling to generate data based on a fixed probability of sampling between variants. The simulation was constructed
to match the original experiment as closely as possible, to aid subsequent data analysis and comparison between experimental and simulated means. Each run of the simulations consisted of 200 agents (i.e. simulated participants), with 50 assigned to each for the four experimental conditions. Each agent generated 16 data points by sampling (with replacement) the two variants of interest, namely the two orders: prenominal and postnominal. The sampling probabilities were set to 0.75 for the majority order and 0.25 for the minority order. Which variant received the higher probability depended on the condition. The variant with a probability of 0.75 always matched the majority variant that participants were exposed to in their specific condition for a given dependent type, whereas the variant with the probability of 0.25 always matched the minority variant. For example, the simulation sampled postnominal order with a probability of 0.75 and prenominal order with a probability or 0.25 for the two postnominal conditions (natural adjective and unnatural genitive) and the reverse for the two prenominal conditions (natural genitive and unnatural adjective). The simulation ran for 10,000 full runs through the experiment, generating a total of 16,000,000 \((16 \times 200 \times 10,000)\) data points.

Based on this data, overall mean output entropy was calculated for each run, as well as for each condition within that run, using the same calculation procedures I employed on the experimental data. As with the experimental data, these values were then transformed to show change in overall entropy (output entropy - input entropy). The same methodology that was used to calculate conditional entropy based on head noun for Experiment 2 was also used on the simulated data. These values for conditional entropy, combined with the values for overall entropy, were used to generate a value for output MI (see section 2.6.2 for details on how these measures were calculated). Again, these were transformed into change in MI (output MI - input MI). These analyses resulted in two sets of five values based on the simulated data for each simulation run. A mean change in entropy for each condition and overall for every run through the simulation (first set), and a mean change in MI based on head noun for each condition and overall for every run through the simulation (second set). The final step of the calculation was to use these values to extract a set of ten simulated mean of means, one mean for each condition and one overall mean for change in entropy, and one mean for each condition and one overall mean for change in MI across the full set of simulated means. A comparison between simulated means of means for change in overall entropy and change in MI compared to their corresponding experimental means can be seen in Figures 2.G.1-2.G.4. As the z-score calculations for Experiment 2 confirmed (see Table 2.5 and 2.7), all experimental values are
significantly different from the simulated means except the change in MI in the natural adjective condition.

Figure 2.G.1: Distribution on simulated means of change in overall entropy across conditions, including visual comparison between simulated mean of means and observed experimental mean for Experiment 2 in Chapter 2. Experimental mean is significantly different from simulated mean of means.
Figure 2.G.2: Distribution on simulated means of change in overall entropy for each condition, including visual comparison between simulated means of means and observed experimental means for Experiment 2 in Chapter 2. Experimental means are significantly different from simulated means of means for each condition.

Figure 2.G.3: Distribution on simulated means of change in MI based on head noun across conditions, including visual comparison between simulated mean of means and observed experimental mean for Experiment 2 in Chapter 2. Experimental mean is significantly different from simulated mean of means.
Figure 2.G.4: Distribution on simulated means of change in MI based on head noun for each condition, including visual comparison between simulated means of means and observed experimental means for Experiment 2 in Chapter 2. Experimental means are significantly different from simulated means of means for each condition except the natural adjective condition.
Chapter 3

Item-specific and system-wide preferences in competition: Effects on noun phrase harmony

3.1 Preface: Chapter 3

Following the results of the experiments in Chapter 2, which found evidence for category-specific biases for postnominal adjective order and prenominal genitive order in a context where participants lacked evidence of the wider language system, but which failed to find any evidence of learning differences between bias-congruent (natural) and bias-incongruent (unnatural) language systems in a small artificial language learning task, this chapter focuses on a slightly different language learning task. This task includes a more complex artificial language system, where participants are trained on orders between both adjectives and nouns as well as orders between genitives and nouns. This differs from the task in Experiment 2 of Chapter 2 where participants were only trained on the order between the noun and one of the two dependent types (either adjectives or genitives). As such, the language that participants are exposed to in this experiment more closely resembles natural language situation where individuals will not be exposed to just one of those dependents in isolation. It was also motivated by the idea that the categories of adjective and genitive might be more clearly activated in participants’ minds in a system which included a contrast between both types of meanings. A stronger activation of the correct item categories would make it more likely that we would observe category-specific biases relating to each dependent, even in a learning context. Finally, while the preference for natural orders in Experiment 2 in Chapter 2 was competing against within-category
harmony (i.e. genitive to genitive or adjective to adjective), here such a preference would compete against cross-category harmony between different dependent types. Given that this is the type of competition we see reflected in typology the experiment in this chapter allows for more direct comparison with the typological data than the results from Experiment 2 in Chapter 2. Just as in the previous experiment, the artificial language learning experiment in this chapter found no learning advantage for natural orders, but again found an overall preference for harmony and reduction in unpredictable variation.

Readers will also notice similarities between the analysis of the experiment in this chapter and Experiment 2 in the previous chapter. I employ the same kind of Monte Carlo simulation to validate the reliability of the entropy and mutual information measures, and calculate values for entropy and mutual information in the same way. Additionally, by combining both adjective and genitive meanings in the same artificial language I was also able to perform a Markov Chain analysis using the probabilities of transition between states (language types) based on a comparison between participants’ training and testing data. This allowed me to derive a stationary distribution of language types (natural, unnatural, harmonic prenominal, and harmonic postnominal) predicted to exist based on the experimental data and compare to the current typological data for these language types. This analysis was not included in the main manuscript below as it was exploratory in nature, but details and results from this analysis can be seen in the chapter appendix, along with details about the modifications made to the Monte Carlo simulation to apply to the more complicated language structure in this experiment.
3.2 Holtz et al. (under review): author contributions

The following paper was submitted to Glossa Psycholinguistics in September 2023. The paper was co-authored with my two supervisors, Jennifer Culbertson and Simon Kirby. The development of the hypotheses, experimental design, and analysis took place during supervision meetings where all three authors were present and contributing. Both co-authors gave feedback during the write-up process. The manuscript has been reformatted to adhere to thesis formatting guidelines. Discrepancies between this paper and any future published version may occur as a result of the review process.
3.3 Abstract

Typological data show a tendency for languages to exhibit harmonic (i.e. consistent) ordering between heads and dependents. Previous experimental work using artificial language learning experiments has shown that learners prefer harmonic patterns. This suggests that the typological trend for harmony may, at least in part, be driven by a cognitive bias. However, it is well-documented that specific categories sometimes contradict this tendency. Here we investigate one such case in the domain of the noun phrase. While many nominal dependents exhibit harmony, adjectives and genitives do not: adjectives tend to follow the noun and genitives tend to precede. Previous experiments have identified the existence of cognitive biases that keep these dependents split across the head noun in contexts where there is no conventional language system in place. In this study we use a silent gesture experiment to examine whether the specific biases that apply to these two dependent types compete with a general preference for harmony in an artificial language learning task. Specifically, we examine whether participants’ learning behaviour is consistent with a preference not just for harmony, but for a non-harmonic order where adjectives follow and genitives precede the noun. What we find instead is that participants’ preference for harmony is not modulated by category-specific biases for prenominal genitives and postnominal adjectives. Thus when participants are learning a conventionalised system, the general bias in favour of harmony overrides category-specific preferences. We discuss the implications of this finding for explanations of typological tendencies which link them to cognitive biases.

Keywords: word order; harmony; regularisation; silent gesture; artificial language learning

3.4 Introduction

Recent research has explored the way that cognitive biases, which favour or disfavour certain linguistic structures, have contributed to the cross-linguistic typological tendency for harmonic word order (Christiansen, 2000; Culbertson, 2012; Culbertson & Kirby, 2016; Wang et al., 2023). This work includes experimental, computational and corpus-based studies examining how these structural biases emerge, and how they affect linguistic behaviour across different age groups and structures. Generally, artificial language learning experiments examining this pattern have found that people tend to prefer harmonic languages to non-harmonic
ones: both adults and children learn, reproduce and regularise (i.e., extend) harmonic patterns rather than non-harmonic ones (Culbertson & Newport, 2017; Culbertson et al., 2012), even if their native languages are non-harmonic for the tested structures (Culbertson, Franck, et al., 2020; Culbertson & Newport, 2015; Wang et al., 2023).

Crucially, the bias to harmonise language structure appears to be particularly active during language learning, in other words, when participants are trained on a language system and then have to reproduce that language at test (Holtz et al., 2023). For example, studies such as Culbertson et al. (2012) and Culbertson, Franck, et al. (2020) use a regularisation design, where the target language participants learn exhibits random variation, here in terms of the word order. In these studies, some participants learn a language that tends toward harmony (i.e., adjectives and numerals mostly on the same side of the noun), and others learn a language that tends toward non-harmony (i.e., adjectives and numerals tend to appear on different sides of the noun). The results show that when the majority word order pattern in these target languages is harmonic participants are more likely to extend the harmonic order and eliminate the variability present in the training data. The idea that the harmony bias is especially active during learning is perhaps unsurprising if we consider the role that harmony plays in providing generalisable rules for a language system. An idealised harmonic language requires encoding of fewer ordering rules than non-harmonic ones (see example in Table 3.4.1), and allows for one high-level rule to dictate head-dependent order across phrase types. In order for the benefits of that type of systematic organisation to be relevant, people must have access to (at least) two different phrase types or dependents in the same context. And so, learning tasks where multiple different phrases are taught to participants naturally should be more likely to reveal the preference for harmonic structures.

<table>
<thead>
<tr>
<th>Harmonic</th>
<th>Non-harmonic</th>
</tr>
</thead>
<tbody>
<tr>
<td>XP $\rightarrow$ X ...</td>
<td>NP $\rightarrow$ N...</td>
</tr>
<tr>
<td>VP $\rightarrow$ ...V</td>
<td>PP $\rightarrow$ P...</td>
</tr>
</tbody>
</table>

Table 3.4.1: Illustrative example of how a harmonic language can be represented with fewer ordering rules than a non-harmonic language.

In contrast, studies in which participants have to improvise linguistic structures in the absence of an existing language system, do not necessarily result in harmonic orders. For example, Culbertson, Schouwstra, and Kirby (2020) asked participants to improvise silent gestures to convey simple objects and their prop-
erties, e.g., ‘these four spotted toothbrushes’. Participants in this study tended to produce non-harmonic gesture orders like THESE FOUR TOOTHBRUSHES SPOTTED, or THESE TOOTHBRUSHES SPOTTED FOUR, where only a subset of the dependents are on the same side of the noun.

In the learning experiments participants only ever had to produce simple noun phrases (i.e. phrases with one dependent at a time), whereas in the improvisation study participants had to produce complex noun phrases with several dependents. The simple noun phrases in the learning experiment may have highlighted the structural parallel between elements across noun phrases in that task, strengthening participants’ tendency to harmonise. Whereas the greater complexity of the meanings that participants had to convey in the improvisational study meant that the structural parallels for harmony were not as clear for the complex noun phrases, reducing participants’ tendency to produce harmonic gesture orders. In addition, there is evidence that improvisation is generally guided by a different type of cognitive bias, which targets ordering preferences for specific linguistic items or categories, rather than guiding the systematic organisation between items (as harmony does). For example, gesture improvisation tasks reveal that different types of verbs favour the use of different basic word orders. Verbs denoting intensional events (e.g. imagine) favour SVO order, and verbs denoting extensional events (e.g. throw) favour SOV order (Motamedi et al., 2022; Schouwstra & de Swart, 2014). Similarly, recent experimental work suggests that adjectives and genitives are also subject to category-specific ordering preferences. When participants have to improvise gestures including a noun and an adjective, they prefer postnominal adjectives (Culbertson, Schouwstra, & Kirby, 2020; Holtz et al., 2023; Jaffan et al., 2020). By contrast, when gestures include nouns and genitives, participants prefer prenominal genitives (Holtz et al., 2023). Interestingly, postnominal adjectives and prenominal genitives are also more common in the typology of both spoken and signed languages (see Table 3.4.2).

Table 3.4.2: Typological counts of the order of nouns with adjective and noun with genitive, in spoken (Spo.) and signed (Sig.) languages (Coons, 2022; Dryer, 2013a, 2013b).\(^1\)

<table>
<thead>
<tr>
<th>Adjective</th>
<th>N (Spo.)</th>
<th>N (Sig.)</th>
<th>Genitive</th>
<th>N (Spo.)</th>
<th>N (Sig.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postnominal</td>
<td>879</td>
<td>16</td>
<td>Postnominal</td>
<td>468</td>
<td>2</td>
</tr>
<tr>
<td>Prenominal</td>
<td>373</td>
<td>10</td>
<td>Prenominal</td>
<td>685</td>
<td>22</td>
</tr>
<tr>
<td>Other</td>
<td>110</td>
<td>15</td>
<td>Other</td>
<td>96</td>
<td>10</td>
</tr>
</tbody>
</table>

\(^1\)Due to the low \(N\) it is hard to evaluate the reliability of this tendency for signed languages.
Typological data on harmony between these two dependents shows evidence of both types of biases at work: while the two harmonic orders are both fairly well-attested, the non-harmonic order which has prenominal genitives and postnominal adjectives is just as common as the most common harmonic order (see Table 3.4.3). This is interesting, because biases which are active during language learning will in principle exert a recurring influence on language structure; individual learning biases of this kind have been shown to be amplified by the process of cultural transmission (Kirby et al., 2007; Thompson et al., 2016). By contrast, category-specific biases—which based on previous experimental work appear to be active mainly during improvisation, more akin to language creation than language learning—presumably have less opportunity to influence typology. And yet, these category-specific biases are visible in the typology. However, two recent studies have found evidence that some category-specific biases may also influence language learning tasks. For example, Motamedi et al. (2022) found better learning/regularisation of language systems where the basic word order for intensional and extensional events matched the order favoured by the category-specific biases mentioned above (SOV for extensional events and SVO for intensional events). Further, Do et al. (2022) also found better learning/regularisation of an artificial language with postnominal adjectives, compared to one with prenominal adjectives (see also Culbertson et al., 2012). Together with the typological data, these studies suggest that the category-specific bias for prenominal genitives and postnominal adjectives are not necessarily restricted to improvisational contexts, and could potentially guide learning behaviour.

We explore this possibility using a gesture-based artificial language learning experiment. Artificial language learning with gestural stimuli is a powerful experimental tool as it allows us to observe cognitive biases that influence participants’ linguistic behaviour, while limiting the effect of participants’ native language (Goldin-Meadow et al., 2008; Motamedi-Mousavi, 2017). Furthermore, it also enables us to manipulate the experimental task to create language contexts that simulate different stages of the evolution of language (Holtz et al., 2023), thereby

The data in this table is based on raw frequencies of languages which are classified rather strictly. Recent research shows that word order tends to exist on more of a gradient than these divisions imply (Levshina et al., 2023), and raw counts also lack controlling factors for language family (Dunn et al., 2011; Hartung et al., 2022). When examining word order correlations while controlling for language family in the sample the results appear mixed, with some studies finding that typological tendencies are highly lineage-specific (Hartung et al., 2022), while others find more evidence for lineage independence (Jäger & Wahle, 2021).

The signed language data in this table was collated by the first author based on available data from Coons (2022). The included data is limited to those languages where a clear order could be defined for both dependent types.
examining which cognitive biases are active at different stages of language evolution. In our study, participants are trained on a gesture system with variable word order, similar to previous experiments targeting regularisation of variation during learning (Culbertson et al., 2012; Ferdinand et al., 2019; Motamedi et al., 2022). This design allows us to compare the degree to which participants harmonise order across dependent types or regularise particular—potentially non-harmonic—orders which match category-specific biases. Importantly, the artificial languages that participants are trained on vary along two variables, namely how much of the system exhibits natural orders for adjectives and genitives, and how harmonic the system is across these two dependents. We compare harmonisation—i.e., the degree to which participants bring dependents into alignment, either before or after the noun—and regularisation—i.e., the degree to which participants reduce variation for specific dependent types—across systems that either align with the individual ordering biases for prenominal genitives and postnominal adjectives, or oppose these biases. If category-specific biases influence learning, then we would predict more regularisation of natural (i.e. bias congruent) word orders compared to unnatural orders, and for harmonisation to be modulated by naturalness.

### 3.5 Experiment

The experiment was a between-subjects design, based on previous work employing regularisation designs in ALL (Culbertson et al., 2012; Ferdinand et al., 2019; Motamedi et al., 2022). In these studies, the variable input presented during training is compared to the output participants produce at test. The prediction is typically that participants are more likely to regularise (or extend) the use of

---

The preregistration for this study can be viewed here [https://osf.io/2us8c/?view_only=05be109ec80b45779d97fe0a9967888f](https://osf.io/2us8c/?view_only=05be109ec80b45779d97fe0a9967888f).

---

Table 3.4.3: Typological counts of the order of noun, genitive and adjective, based on spoken (Spo.) language typological data (Dryer, 2013a, 2013b) and signed (Sig.) language typological data (Coons, 2022). Harmonic patterns are common, but so is a particular non-harmonic ordering, with postnominal adjectives and prenominal genitives (bolded).

<table>
<thead>
<tr>
<th>Order</th>
<th>N (Spo.)</th>
<th>N (Sig.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-Adj / N-Gen</td>
<td>342</td>
<td>-</td>
</tr>
<tr>
<td>Adj-N / Gen-N</td>
<td>232</td>
<td>8</td>
</tr>
<tr>
<td>N-Adj / Gen-N</td>
<td>342</td>
<td>9</td>
</tr>
<tr>
<td>Adj-N / N-Gen</td>
<td>65</td>
<td>-</td>
</tr>
</tbody>
</table>
preferred variants, compared to dis-preferred variants (here dis-preferred orders).

All participants in our study were exposed to a majority and a minority gesture order for both adjective and genitive meanings during training. The majority order appeared in 75% of training trials and the minority order in the remaining 25% of training trials. The main manipulation between the four experimental conditions was in whether the majority order was prenominal or postnominal for each dependent type (see Table 3.5.1 for condition majority/minority orders). Two of the conditions were majority harmonic (both dependents tended to be prenominal, or both tended to be postnominal). Two of the conditions were majority non-harmonic (dependents tended to come on opposite sides of the noun). Of the latter, one (natural condition) matched the hypothesised category-specific biases, with adjectives tending to be post-nominal and genitives tending to be prenominal, and other (unnatural condition) did not.

Table 3.5.1: Percentage of prenominal and postnominal gesture orders for each dependent type in the training data across the four experimental conditions.

<table>
<thead>
<tr>
<th>Condition Order</th>
<th>Adjective</th>
<th>Genitive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>Pre: 25%</td>
<td>Post: 75%</td>
</tr>
<tr>
<td>Genitive</td>
<td>Pre: 75%</td>
<td>Post: 25%</td>
</tr>
<tr>
<td>Natural (non-harmonic)</td>
<td>25%</td>
<td>75%</td>
</tr>
<tr>
<td>Unnatural (non-harmonic)</td>
<td>75%</td>
<td>25%</td>
</tr>
<tr>
<td>Prenominal (harmonic)</td>
<td>75%</td>
<td>25%</td>
</tr>
<tr>
<td>Postnominal (harmonic)</td>
<td>25%</td>
<td>75%</td>
</tr>
</tbody>
</table>

3.5.1 Methods

Materials

The experiment was built in the JavaScript library jsPsych (de Leeuw et al., 2023) and ran in participants’ web browsers. The experiment had two main sets of stimuli. First, a collection of grayscale digital drawings depicting instances of item ownership (genitives) or items with different patterns (adjectives). The total set comprised 16 images, with every combination of the four items (book, cup, hat, and scarf), the two possessors (vampire and cyclops), and two patterns (spots and stripes). Examples of the images can be seen in Figure 3.5.1.

Paired with each image there were two corresponding gesture videos. The videos showed a model gesturer producing an expression signifying the meanings of the images using a sequence of two gestures (see Figure 3.5.2 for example stills from one of the gesture videos). One gesture represented the head noun and one the adjective/genitive dependent. Each pair of videos differed only in the order
Figure 3.5.1: Sample stimuli from adjective conditions (left) and genitive condition (right).

(a) Gesture for noun *book*.

(b) Gesture for adjective *striped*.

Figure 3.5.2: Stills of gestures used to denote the meaning *striped book*.

of these two gestures – in one video the sequence was head initial and the in the other it was head final. Each meaning component was denoted using a two-handed gesture and each sequence ended simultaneously in a neutral position with both hands clasped just below waist-height. All videos were 4.389 seconds long.

**Procedure**

Participants were randomly assigned to one of the four conditions seen in Table 3.5.1, and a pseudo-randomised stimuli set consisting of four meanings.\footnote{This stimuli set was comparable in size to previous studies using similar methodology e.g. Motamedi et al. (2022) where Experiment 2a had two meanings and Experiment 2b had six meanings in the training set.} The complete stimuli set consisted of two nouns, one randomly selected from the set of “worn” items (*hat* and *scarf*) and one randomly selected from the set of “held” items (*cup* and *book*). These both appeared with one of the genitives and one of the adjectives respectively. At the start of the experiment participants were told that they would learn to express ownership and describe items in a made-up sign
language. They were then familiarised with the stimuli by being shown two trials including 2x2 grids, one of adjective meanings and one of genitive meanings, like those in Figure 3.5.1.

During the training phase participants saw these same grids, but with one image highlighted by a red square. Under the images a gesture video associated with the highlighted meaning was played. Each of the four images in the assigned stimuli set acted as targets eight times each, making 32 training trials. The proportion of prenominal versus postnominal gesture orders in the training videos varied depending on the condition, but 6/8 times the gesture video was the majority order for that condition, and twice it was the minority order. The training phase progressed automatically and participants were instructed sit back and watch the videos.

In the testing phase, participants saw the same image grids as during training with the target image highlighted as during training. However, this time both possible gesture videos (prenominal and postnominal gesture order) were looping in aligned positions beneath the images. Participants were instructed to “click on the corresponding gesture video” that was like the ones they had seen during training. Between each of the 32 test trials (eight per meaning) participants also clicked a centred “next” button to reset the mouse. Following the testing phase, participants provided English translations for all four target meanings, and gave details regarding their proficiency in signed and spoken languages.\(^5\)

**Participants**

211 participants were recruited through the online crowdsourcing platform Prolific. We prescreened participants for the following criteria: 95% previous task approval rate, reported English as their first language, and no previous participation in our experiments/pilots. Participants were paid the equivalent of £8.91 per hour.\(^6\)

\(^5\)Details regarding spoken and sign-language proficiency for participants can be found in the OSF repository linked in the data availability statement. Translation data is available in the survey responses.

\(^6\)Following preregistered exclusion criteria we excluded 8 participants who reported proficiency in a signed language, as well as 3 participants who always pressed the same button during testing. The data in the analysis comes from 50 participants in the natural condition, 53 in the unnatural condition, 46 in the majority prenominal condition, and 51 in the majority postnominal condition.
3.5.2 Results

The main predictions for this experiment were as follows. We predicted that participants would generally learn the orders they saw in training (i.e. choose the majority gesture order from the training data). However, we predicted that natural orders would be more likely to be chosen than unnatural orders (i.e., prenominal orders for genitive meanings and postnominal orders for adjective meanings) across conditions, and that naturalness would modulate learning such that natural orders would be learned better than unnatural orders. Based on previous evidence (e.g. Ferdinand et al., 2019), we also predicted that participants’ output (as reflected by their video choices at test) would be less variable than the input—i.e., that participants would regularise. They could regularise either within the context of one dependent type, by choosing one order for each dependent type, or across both dependent types, by selecting one order across the whole system. The latter type of regularisation would indicate a bias for harmony.

Figure 3.5.3 shows the overall distribution of postnominal orders that participants selected across all conditions.

![Figure 3.5.3](image)

Figure 3.5.3: Proportion of postnominal orders selected by each participant for both dependent types across all conditions. Each small point corresponds to a single participant, coloured by the condition they are in. The four larger shapes represent the input proportions in each condition. Points in the corners are completely regular for both modifiers. The lower left and upper right corners correspond to harmonic orders, whereas the upper left corner corresponds to natural orders for both modifiers and the lower right corner corresponds to fully unnatural orders for both modifiers.
Learning

Figure 3.5.4 shows the proportion of majority orders selected during the test phase by participants in each condition, as well as the overall grand mean. To examine learning patterns across conditions we ran a mixed effects logistic regression model with *majority order* as the binary outcome variable (1 indicating a match between majority training order and order selected at test) and with *condition* and *dependent type*, as well as their interaction, as fixed effects. Both fixed effects were deviation coded. The model also included a by-participant random intercept, and a random slope for dependent type. The model showed a significant positive intercept ($\beta = 1.34$, $SE = 0.12$, $z = 11.24$, $p < 0.001$), indicating that participants across all conditions, on average, chose the majority order at a rate above chance. *Condition* was also significant, as participants in the majority prenominal condition chose the majority order more often, compared to the grand mean ($\beta = 0.64$, $SE = 0.21$, $z = 3.05$, $p < 0.01$). Additionally, participants tended to select the majority order less often for genitive meanings ($\beta = -0.74$, $SE = 0.18$, $z = -4.24$, $p < 0.001$), except in the natural condition, where they selected more majority orders for genitives ($\beta = 0.99$, $SE = 0.29$, $z = 3.40$, $p < 0.001$), compared to the grand mean.

![Graph showing proportion of test trials where participants selected the majority input order for each condition and dependent type.](image)

Figure 3.5.4: Proportion of test trials where participants selected the majority input order for each condition and dependent type. Right-hand facet shows grand mean. Participants tended to reproduce the majority orders from their training. Error bars show bootstrapped 95% CIs.
Naturalness

Figure 3.5.5 shows the proportion of natural orders (postnominal for adjectives and prenominal for genitives) selected at test by participants in each condition, as well as the overall grand mean. To examine whether selection of natural orders differed by condition, we implemented a logistic mixed effects model with natural as the binary outcome variable (1 indicating a match between the predicted natural order and selected order, 0 meaning a mismatch). The fixed and random effect structure was the same as for the learning model described above. As Figure 3.5.5 suggests, the intercept was non-significant. However, there was an effect of condition, such that natural orders were more likely to be selected in the natural ($\beta = 1.14, SE = 0.18, z = 6.49, p < 0.001$) and majority postnominal conditions ($\beta = 0.59, SE = 0.17, z = 3.36, p < 0.001$), compared to the grand mean. Conversely, there was a negative effect of condition on naturalness in the unnatural ($\beta = -1.24, SE = 0.17, z = -7.11, p < 0.001$) and majority prenominal conditions ($\beta = -0.88, SE = 0.18, z = -2.62, p < 0.01$). Finally, the interaction between condition and dependent type was significant for some conditions. In the majority prenominal condition, participants selected the natural order more often for genitive meanings ($\beta = 3.03, SE = 0.36, z = 8.40, p < 0.001$). Whereas in the majority postnominal condition, participants selected the natural order less often for genitives ($\beta = -2.71, SE = 0.34, z = -7.94, p < 0.001$).

![Proportion of test trials where participants selected the natural order (i.e., prenominal for genitives, postnominal for adjectives) for each condition and dependent type. Right-hand facet shows the grand mean. There was no overall preference for natural orders. Error bars show bootstrapped 95% CIs.](image)

Figure 3.5.5: Proportion of test trials where participants selected the natural order (i.e., prenominal for genitives, postnominal for adjectives) for each condition and dependent type. Right-hand facet shows the grand mean. There was no overall preference for natural orders. Error bars show bootstrapped 95% CIs.
Harmony

A fully harmonic language refers to a system in which all meanings are expressed with the same gesture order. As such, harmony constitutes a reduction in overall system variation. We quantify an increase in harmony as an overall reduction in entropy, following Ferdinand et al. (2019) and Motamedi et al. (2021). The entropy ($H$) of a system is defined as:

$$H(V) = -\sum_{v_i \in V} p(v_i) \log_2 p(v_i)$$

where, ($V$) refers to the two gesture variants (prenominal and postnominal orders). Both the natural and unnatural training data had an entropy value of 1, as participants were exposed to a 50/50 split of both variants (orders). The remaining two conditions have a training entropy of 0.8112781. We measured the change in entropy between these input values and the orders participants selected (output entropy - input entropy) to see whether participants tended to reduce variation by harmonising the system. Figure 3.5.6 shows the mean change in entropy for each condition as well as the overall mean change.

![Figure 3.5.6: Mean change in entropy between input and output, showing participants tend to harmonise, except in the postnominal condition. Right-hand facet shows grand mean change in entropy. Error bars show bootstrapped 95% CIs.](image)

To evaluate whether these changes are reliably greater than zero we calculated bootstrapped confidence intervals around the reported mean entropy changes for
These were generated using the ‘boot’ package in R and based on 10,000 samples. Our findings suggest that all but one of the conditions—majority postnominal—show a reliable reduction in entropy. These results were verified by a Monte Carlo simulation, where the probability of sampling prenominal and postnominal order was set to the training proportions for each condition. Based on 10,000 simulations, mean changes in entropy were extracted across all conditions, as well as individual means for each condition. Z-scores were calculated based on simulated and experimental means, with results showing that change in entropy was reliable overall, as well as for each condition except the majority postnominal condition (see Table 3.5.2 for z-scores). Furthermore, confidence intervals around differences between conditions reveals that the majority prenominal condition is reliably different from all other conditions (nat - majPre: $x_a - x_b = 0.28$, lower CI = 0.16, upper CI = 0.41; unnat - majPre: $x_a - x_b = 0.26$, lower CI = 0.14, upper CI = 0.38; majPre - majPost: $x_a - x_b = -0.17$, lower CI = -0.31, upper CI = -0.03). Similarly, the majority postnominal condition is also reliably different from the natural condition ($x_a - x_b = 0.12$, lower CI = 0.02, upper CI = 0.23).

Table 3.5.2: Experimental means, simulated means and resulting z-scores for change in overall entropy. Z-scores show that experimental means are reliably different from the simulated means (indicating that entropy dropped significantly) in all conditions except the majority postnominal condition.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Exp. mean</th>
<th>Sim. mean</th>
<th>z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>-0.102</td>
<td>-0.020</td>
<td>-12.691</td>
</tr>
<tr>
<td>nat</td>
<td>-0.092</td>
<td>-0.017</td>
<td>-22.043</td>
</tr>
<tr>
<td>unnat</td>
<td>-0.119</td>
<td>-0.017</td>
<td>-29.555</td>
</tr>
<tr>
<td>majPre</td>
<td>-0.186</td>
<td>-0.023</td>
<td>-9.175</td>
</tr>
<tr>
<td>majPost</td>
<td>-0.019</td>
<td>-0.023</td>
<td>0.249</td>
</tr>
</tbody>
</table>

Exploratory analysis: Conditional entropy and mutual information

In addition to harmony, participants may have reduced the unpredictable variation in the training data by introducing conditioning contexts for the variants (orders). There are two potential conditioning contexts, namely dependent type or head noun. To examine if either of these strategies were used by participants we first calculated two types of conditional entropy for each participant based on

---

7Linear models are not reported for this data as the distribution of entropy in our study is necessarily non-normal.
this formula:

\[ H(V|C) = - \sum_{c_j \in C} p(c_j) \sum_{v_i \in V} p(v_i|c_j) \log_2 p(v_i|c_j) \]

The contexts (C) here stands for dependent type in one calculation, and head noun in the other.\(^8\) These calculations of conditional entropy also capture reduction in variability due to harmony (becoming more consistent across the whole system will also result in higher consistency within sub-contexts). To disentangle harmony from these other regularisation strategies we used the extracted conditional entropy values to calculate Mutual Information of selected gesture order for both dependent type

\[ \text{Mutual Information (MI)} = \text{overall entropy} - \text{conditional entropy (dependent)} \]

and head noun

\[ \text{Mutual Information (MI)} = \text{overall entropy} - \text{conditional entropy (noun)} \]

A system with an MI value of 1 would mean that variants were perfectly conditioned on either the two dependent types or the head noun, whereas MI of 0 indicates that the variability within each context mirrors the whole system. Figure 3.5.7 shows overall mean MI change (i.e. output MI - input MI) across conditions for dependent type, as well as individual condition means. Figure 3.5.8 shows overall mean MI change across conditions for head noun, as well as individual condition means. Confidence intervals were generated in the same way as for our harmony measure. Based on the simulations, means for both types of conditional entropy were extracted and z-scores calculated. Table 3.5.3 shows z-scores based on dependent type and Table 3.5.4 shows z-scores based on head noun. Increase in MI is significant across all conditions for both measures.\(^9\)

\(^8\)The preregistration for this study did not include head noun as a context for conditional entropy, it was included here after noting that lexical contexts is a common strategy for conditioning unpredictable variation (Smith & Wonnacott, 2010).

\(^9\)Note that this MI calculation was not included in the original preregistration and is therefore reported as an exploratory analysis.
Figure 3.5.7: Mean change in MI based on dependent type, showing reliable change in MI across all conditions except the natural condition. Note that the significance of this measure differs between that presented in this figure and the $z$-scores comparing simulated and experimental means. Right-hand facet shows grand mean change in MI. Error bars show bootstrapped 95% CIs.

3.6 Discussion

This experiment examined whether the previously observed preference for word order harmony during learning would be disrupted by the presence of category-specific ordering preferences. Participants were taught a system with variable word order for nouns with adjectives and genitives which either tended toward harmony (e.g., Noun-Genitive, Noun-Adjective), or tended toward non-harmony (e.g., Noun-Genitive, Adjective-Noun). Crucially, one of the non-harmonic orders aligns with category-specific biases for these dependent types previously found in improvisation tasks (Genitive-Noun, Noun-Adjective). The results of the experiment indicate that a general tendency toward harmony persists in this task (although weakly), and is not clearly modulated by naturalness: a match between majority training order and preferred order for adjectives (postnominal) and genitives (prenominal) did not result in more consistent regularisation of that non-harmonic pattern. Additionally, participants also showed some tendency to condition the word order they used based on the meaning they were trying to convey. This finding is broadly consistent with previous work, showing that people tend to reduce variation by either eliminating competing variants (here harmonising gesture orders) or conditioning the variants on some aspect of the system (often related to meaning; see Smith & Wonnacott, 2010; Smith et al., 2017).
Figure 3.5.8: Mean change in MI based on head nouns, showing slight increase in MI across all conditions. Right-hand facet shows grand mean change in MI. Error bars show bootstrapped 95% CIs.

Table 3.5.3: Experimental means, simulated means and resulting z-scores for change in MI based on dependent type. Z-scores show that experimental means are reliably different from the simulated means (indicating that MI increased significantly) in all conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Exp. mean</th>
<th>Sim. mean</th>
<th>z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.114</td>
<td>0.028</td>
<td>12.805</td>
</tr>
<tr>
<td>nat</td>
<td>0.083</td>
<td>0.031</td>
<td>2.800</td>
</tr>
<tr>
<td>unnat</td>
<td>0.139</td>
<td>0.031</td>
<td>5.834</td>
</tr>
<tr>
<td>majPre</td>
<td>0.094</td>
<td>0.025</td>
<td>13.835</td>
</tr>
<tr>
<td>majPost</td>
<td>0.139</td>
<td>0.025</td>
<td>22.510</td>
</tr>
</tbody>
</table>

While there was an overall increase in harmony in participants’ gesture order selections in the testing phase, this was not consistent across all conditions. Participants in the majority prenominal condition harmonised more readily than participants in all the other conditions. In contrast, participants in the majority postnominal condition showed no clear tendency to harmonise. One possible explanation for the difference between these conditions is the way the majority training orders align with the word orders in participants’ native language. All participants in our experiment reported English as their first language, meaning that the orders in the majority prenominal condition were the most similar to English for these dependent types (mostly prenominal genitives and prenominal adjectives), whereas the majority orders in the majority postnominal condition...
Table 3.5.4: Experimental means, simulated means and resulting z-scores for change in MI based on head noun. Z-scores show that experimental means are reliably different from the simulated means (indicating that MI increased significantly) in all conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Exp. mean</th>
<th>Sim. mean</th>
<th>z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.092</td>
<td>0.021</td>
<td>32.639</td>
</tr>
<tr>
<td>nat</td>
<td>0.131</td>
<td>0.018</td>
<td>31.946</td>
</tr>
<tr>
<td>unnat</td>
<td>0.071</td>
<td>0.018</td>
<td>14.949</td>
</tr>
<tr>
<td>majPre</td>
<td>0.096</td>
<td>0.025</td>
<td>14.129</td>
</tr>
<tr>
<td>majPost</td>
<td>0.072</td>
<td>0.025</td>
<td>9.436</td>
</tr>
</tbody>
</table>

are the opposite of the English word orders. This suggests that, even though previous research has found that using silent gesture limits structural transfer from participants’ native language (Goldin-Meadow et al., 2008), this methodology does not completely eliminate this type of transfer (Jaffan et al., 2020).

The failure of naturalness to modulate the preference for harmony contrasts with the results of two previous studies. Do et al. (2022) examined regularisation behaviour for prenominal and postnominal adjectives in a gesture learning task. They found that participants were more likely to extend a majority gesture order if that order was postnominal, rather than prenominal. Despite the parallels between Do et al.’s (2022) study and ours, there are also some notable differences. The meanings in their study include nominals modified by adjectives, but also verb-like meanings (e.g., wave). In addition, that study uses videos rather than still images for the prompt for meaning. It is possible that the visual parallels between the movements in the stimuli videos and the gestural videos activated the postnominal preference more strongly in their task than the ones we employed here. Little direct research has compared the effect of stimuli medium (videos vs images) on the behaviour of participants in silent gesture experiments, but at least one study has found that even small variations in stimuli can affect participants’ gesture preferences (Kirton et al., 2021).

Motamedi et al. (2022) observed that, in a similar task to the one used in this study, category-specific biases for using SOV order for extensional events and SVO order for intentional events influenced regularisation behaviour. Systems which followed the preferred order patterns in their study were more likely to be systematically regularised. The difference between that study and ours may be due to weaker category-specific ordering preferences for adjectives and genitives than for event-type conditioned ordering preferences. Some tentative support for this with regards to adjectives can be seen in the learning behaviour of participants.
in the harmony study by Culbertson et al. (2012). There, participants who were trained on a majority non-harmonic system with postnominal adjectives showed better learning than participants trained on a non-harmonic system with prenominal adjectives (Culbertson et al., 2012). The authors proposed that this pattern might be driven by a specific bias targeting adjective ordering, as we also suggest here. This postnominal adjective preference did not replicate consistently in all the subsequent follow-ups targeting different populations (Culbertson, Franck, et al., 2020; Culbertson & Newport, 2015, 2017), suggesting that the effect of this category-specific bias is difficult to capture in learning tasks. However, it is worth noting that the basic word order bias in Motamedi et al. (2022), originally reported by Schouwstra and de Swart (2014), is not regularly observed in the typology of spoken languages (although see Flaherty et al., 2018; Napoli et al., 2017, for evidence in sign languages). Whereas the ordering preferences for adjectives and genitives do seem to be observable in the typology of both spoken and signed languages (Coons, 2022; but see footnote 1).

The distinction between these two patterns (the one observed for basic word order and the one for noun phrase dependents) could be caused by differing communicative pressures to regularise/harmonise the language system over time. For example, during early stages of language evolution it is possible that languages which had some combination of SVO/SOV word order were widespread (Maslova, 2000; Newmeyer, 2000), and that event-type conditioned variability between these orders within a language was common as a result of the category-specific biases which are active during language creation (Motamedi et al., 2022). Similarly, the favoured order of postnominal adjectives and prenominal genitives may have been present for similar reasons. To explain the difference in the typology for these two patterns one explanation might be that the communicative pressure to converge on a shared, consistent basic word order across all verb types is stronger than the harmony bias present during learning for noun phrase dependents. This could then lead to the present-day typological distribution where event-type conditioned basic word order is rare (Flaherty et al., 2018; Napoli et al., 2017), but non-harmonic noun phrase order is still fairly common (Dryer, 2013a, 2013b).

In combination, the results of previous research and those presented in this study reveal a complicated view of the interaction between system-wide and category-specific biases and their effect on typological word order patterns. It seems as though category-specific biases are present most strongly in improvisation tasks, but that the bias for postnominal adjectives is sometimes present in learning contexts as well. The lack of a naturalness effect in our study could be due to the fact that this task activated both system-wide and category-specific
biases in a limited way, especially given that the observed harmony preference was fairly weak. There are a number of reasons for why this could be. For example, the stimuli might not have invoked the intended categories (adjective, genitive and noun) strongly enough for the related biases to be sufficiently activated. Furthermore, the Mutual Information calculations revealed that participants also employed strategies other than harmonisation to reduce unpredictable variation, such as conditioning the gesture orders based on dependent types and lexical heads. These other strategies may have been especially salient to participants as there were two dependents and two lexical heads in the stimuli sets, which made the strategy of pairing these with one of the two orders a more obvious option.

3.7 Conclusion

The prevalence of typological regularities across the world’s languages suggests that some language features are favoured by our cognitive system. Previous research has identified a number of possible cognitive biases that give rise to specific cross-linguistic patterns, such as the system-wide bias favouring harmonic word order during learning tasks. However, exceptions to this harmonic tendency suggest that other, more category-specific, biases can compete with this general preference for harmony, causing some noun phrase dependents to deviate from the harmonic tendency. These category-specific biases have mainly been observed in improvisation tasks, yet some recent evidence suggests that they may also influence learning behaviour. We tested this possibility for two noun phrase dependents which deviate from the general harmonic trend, namely adjectives and genitives. Using a gestural artificial language learning experiment with a regularisation design we examined if natural systems (i.e. systems with the favoured orders for these dependents) were learned better than systems with unnatural orders. Our experiment found no better learning for natural systems, nor a modulating effect of naturalness on participants’ harmonic or regularisation tendencies. Consequently, the residual pattern we see in the typology does not seem to be the result of recurring application of competition between category-specific and system-wide biases in successive generations of language learners. Instead, we mainly found evidence of different strategies for reducing unpredictable variation, by either extending the use of one order across the whole system (harmonisation) or by conditioning the use of different word orders on certain categories or lexical items in the data. The locus of competition between these types of biases thus remains unclear, and future research is needed to establish under what conditions
category-specific biases influence linguistic behaviour and if this influence extends beyond contexts of language improvisation.

Data Availability Statement

The data collected for this study, along with all analysis and experimental materials, is openly available on OSF and can be found here https://osf.io/n7vhs/?view_only=f42f7a469dd34f6ebd1d3a22bee6cc08.
References


Do, M. L., Kirby, S., & Goldin-Meadow, S. (2022). Regularization of word order in the verb phrase differs from the noun phrase: Evidence from an online silent gesture perception paradigm. In J. Culbertson, A. Perfors,

Dryer, M. S. (2013a). Order of adjective and noun. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology.

Dryer, M. S. (2013b). Order of genitive and noun. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology.


Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B: Biological


Appendix

3.A Outline

This appendix includes further information about additional data collected in the experiment from Chapter 3 as well as an additional analysis that was conducted for this experiment. Sections 3.B and 3.C give details about how informed consent was obtained from participants, and the instructions they saw during the experiment. Following this, 3.D shows the full table where CIs for change in overall entropy are compared between conditions. 3.E gives details about the spoken languages that participants reported knowing when they answered the questionnaire, and section 3.F shows a summary of classifications given for the translation data across the full experiment. Section 3.G describes the set-up and results of the Monte Carlo simulation that was used to evaluate the reliability of the changes in entropy and Mutual Information that were observed in the experiment. Finally, Section 3.H describes the procedure and results of the Markov chain analysis that I conducted based on the experimental results.

3.B Consent info

The experiment included a web-page where participants could indicate their consent to take part in the study by ticking a box and pressing a button to progress the experiment. This page also included the information sheet, with specific information about data storage, right to withdraw etc. Below you will find all the text included on this web-page.
Study title: Gesture Study

Principal Investigators: Jennifer Culbertson, Simon Kirby
Researcher collecting data: Annie Holtz

What is this document? This document explains what kind of study we're doing, what your rights are, and what will be done with your data. You should print this page for your records.

Nature of the study. You are invited to participate in a study that involves observing videos showing a person using gesture in a made-up sign language to convey information. You will see stimulus pictures accompanied by two videos showing gestures, and be asked to make a choice between them. Once you finish, we may have also some questions about your experience (e.g., age, language background). Your responses will be recorded. Your session should last for up to 15 minutes. You will be given full instructions shortly.

Compensation. You will be paid £2.23 for your participation in this study.

Risks and benefits. There are no known risks to participation in this study. Other than the payment mentioned, there are no tangible benefits to you, however you will be contributing to our knowledge about language.

Confidentiality and use of data. All the information we collect during the course of the research will be processed in accordance with Data Protection Law. In order to safeguard your privacy, we will not share personal information (like your name) with anyone outside the research team. Your data will be referred to by a unique participant number rather than by name. Please note that we will temporarily collect your worker ID to prevent repeat participation, however we will never share this information with anyone outside the research team and it will be deleted two weeks after data collection. We will store any personal data (i.e., worker ID) using the University of Edinburgh's secure encrypted storage service. The anonymised data collected during this study will be used for research purposes.

What are my data protection rights? The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance with Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Voluntary participation and right to withdraw. Your participation is voluntary, and you may withdraw from the study at any time and for any reason. If you withdraw from the study during data gathering, we will delete your data and there is no penalty or loss of benefits to which you are otherwise entitled. If you wish to withdraw after you have completed the study we will only be able to delete your data, or to remove it from inclusion in published results for two weeks after the date of your participation. Once again, withdrawing at this stage will not cause any penalty or loss of benefits to which you are otherwise entitled.

If you have any questions about what you've just read, please feel free to ask, or contact us later. You can contact us by email at . This project has been approved by PPLS Ethics committee. If you have questions or comments regarding your rights as a participant, they can be contacted at 0131 650 4020 or ppls.ethics@ed.ac.uk.

By taking part in the experiment, you consent to the following:
1. I agree to participate in this study.
2. I confirm that I have read and understood how my data will be stored and used.
3. I understand that I have the right to terminate this session at any point. If I choose to withdraw after completing the study, and request for my data to be removed within two weeks of my participation, my data will be deleted.

To continue, tick the checkbox below and click "Start Experiment".

- [ ] I agree to take part in this study.

Start Experiment
3.C Instructions

The instructions were identical across all four conditions in the experiment. Below are all the instructions from each segment of the experiment.

Contrast trial instructions: Instructions provided prior to the contrast/familiarisation trials across all conditions.

*In the first part of the experiment, you will observe gestures being produced which correspond to the meaning of different images. The gestures show how you express ownership of an item or how you describe an item in a made-up sign language. On the next two screens you will see examples of the kinds of images you will see in the rest of the experiment. The images vary by the item they show, the owner and the pattern. Study the images and then continue to the next screen.*

Training instructions: Instructions provided before the training phase across all conditions.

*In the first part of the experiment, one of the images in the grid will be highlighted and underneath the images you will see a video of a person using gestures to convey the meaning of the highlighted image. These kinds of images, along with their accompanying gestures, will be shown to you several times. Just sit back and watch carefully.*

Testing instructions: Instructions provided before the testing phase across all conditions.

*In this second part of the experiment, you will be shown the same collection of images again. For each target image, please click on the corresponding gesture video like you saw in the first part. This will repeat several times for each picture.*

Translation instructions: Instructions provided before the translation trials across all conditions.

*You will now be asked to translate some of these gesture sequences into English, please type your answer in the box for each question.*

Questionnaire prompts: Instructions and prompts presented during the questionnaire across all conditions.
Please answer the following questions:

1. Do you have any knowledge of a sign language? (If yes, please state which).

2. What spoken languages do you know and to what proficiency? Please indicate proficiency on a scale of 1-10 (e.g. English - 10, Tamil - 8, French - 4).

### 3.D Entropy CIs comparison

Table 3.D.1: Comparison of overall mean entropy change between conditions for the experiment in Chapter 3. Includes 95% bootstrapped CIs around each mean. Differences across conditions are reliably non-zero for some conditions. The natural condition is reliably different from the majority prenominal and majority postnominal condition. Similarly, the unnatural condition is reliably different from the majority prenominal condition. And the majority prenominal condition is reliably different from the majority postnominal condition.

<table>
<thead>
<tr>
<th>Condition comparison</th>
<th>$\bar{x}_a - \bar{x}_b$</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>nat. – unnat.</td>
<td>0.03</td>
<td>-0.05</td>
<td>0.10</td>
</tr>
<tr>
<td>nat. – maj.prenom.</td>
<td>0.28</td>
<td>0.16</td>
<td>0.41</td>
</tr>
<tr>
<td>nat. – maj.postnom</td>
<td>0.12</td>
<td>0.02</td>
<td>0.23</td>
</tr>
<tr>
<td>unnat. – maj.prenom.</td>
<td>0.26</td>
<td>0.14</td>
<td>0.38</td>
</tr>
<tr>
<td>unnat. – maj.postnom</td>
<td>0.09</td>
<td>-0.003</td>
<td>0.20</td>
</tr>
<tr>
<td>maj.prenom. – maj.postnom.</td>
<td>-0.17</td>
<td>-0.30</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

### 3.E Language data

Data in Table 3.E.1 shows the spoken languages that participants in the experiment reported having any knowledge of. As participants knew several languages the counts here exceeds the total number of participants. Languages are sorted from most frequent to least frequent, with alphabetical ordering for languages with the same number of speakers.
Table 3.E.1: Languages (other than English) spoken by participants whose data was included in the experiment in Chapter 3.

<table>
<thead>
<tr>
<th>Languages</th>
<th>N participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>41</td>
</tr>
<tr>
<td>Spanish</td>
<td>32</td>
</tr>
<tr>
<td>German</td>
<td>13</td>
</tr>
<tr>
<td>Italian</td>
<td>5</td>
</tr>
<tr>
<td>Cantonese</td>
<td>3</td>
</tr>
<tr>
<td>Japanese</td>
<td>3</td>
</tr>
<tr>
<td>Afrikaans</td>
<td>2</td>
</tr>
<tr>
<td>Farsi</td>
<td>2</td>
</tr>
<tr>
<td>Irish</td>
<td>2</td>
</tr>
<tr>
<td>Korean</td>
<td>2</td>
</tr>
<tr>
<td>Mandarin</td>
<td>2</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
</tr>
<tr>
<td>Setswana</td>
<td>2</td>
</tr>
<tr>
<td>Arabic</td>
<td>1</td>
</tr>
<tr>
<td>Armenian</td>
<td>1</td>
</tr>
<tr>
<td>Danish</td>
<td>1</td>
</tr>
<tr>
<td>Hakka</td>
<td>1</td>
</tr>
<tr>
<td>Hebrew</td>
<td>1</td>
</tr>
<tr>
<td>Hindi</td>
<td>1</td>
</tr>
<tr>
<td>Latin</td>
<td>1</td>
</tr>
<tr>
<td>Polish</td>
<td>1</td>
</tr>
<tr>
<td>Punjabi</td>
<td>1</td>
</tr>
<tr>
<td>Sepedi</td>
<td>1</td>
</tr>
<tr>
<td>Sesotho</td>
<td>1</td>
</tr>
<tr>
<td>Swahili</td>
<td>1</td>
</tr>
<tr>
<td>Telugu</td>
<td>1</td>
</tr>
<tr>
<td>Welsh</td>
<td>1</td>
</tr>
<tr>
<td>Zulu</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 3.F.1: Classification of English translations given for target meanings for
the experiment in Chapter 2. Note that each participant provided four transla-
tions, two for adjective meanings and two for genitive meanings. This makes for
a total of 800 translations across all participants.

<table>
<thead>
<tr>
<th>Class</th>
<th>N translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>228</td>
</tr>
<tr>
<td>Genitive</td>
<td>19</td>
</tr>
<tr>
<td>Preposition</td>
<td>191</td>
</tr>
<tr>
<td>Verb phrase</td>
<td>111</td>
</tr>
<tr>
<td>Other</td>
<td>251</td>
</tr>
<tr>
<td>Total</td>
<td>800</td>
</tr>
</tbody>
</table>

3.F  Translation data

3.G  Monte Carlo simulation

Just as in 2.G I performed a Monte Carlo simulation to validate the reliability of
the changes in overall entropy and Mutual Information that were calculated based
on the experimental data from Chapter 3. This simulation was used to compare
the observed experimental values for change in overall entropy and change in
Mutual Information (MI) based on both head noun and dependent type, to a
theoretical baseline of sampling based on input probability.

The simulation was adapted from the one used in Experiment 2 of Chapter 2
to reflect that the experiment in this chapter included both adjective and genitive
meanings in the training data. The simulation ran for 10,000 runs through the ex-
periment, with 50 agents (i.e. simulated participants) assigned to each condition.
Since the testing phase in this experiment involved 32 trials, each agent in the
simulation sampled between the two variants (prenominal or postnominal order)
32 times, generating 64,000,000 (32 * 200 * 10,000) data points. The sampling
probabilities were set to 0.75 for the majority order and 0.25 for the minority
order. Which variant received the higher probability depended on the condition.
The variant with a probability of 0.75 always matched the majority variant that
participants were exposed to in their specific condition for a given dependent
type, whereas the variant with the probability of 0.25 always matched the mi-
nority variant. For example, the simulation sampled postnominal order with a
probability of 0.75 and prenominal order with a probability or 0.25 for adjective
meanings in the natural condition, and sampled with the reverse probability for
genitive meanings in the same condition.

The same procedure was used for both the experimental and simulated data to calculate mean change in overall entropy and mean change in MI based on head nouns for each condition and across conditions for each run of the simulation. Additionally, the same calculation of change in MI was also conducted but with dependent type as the conditioning context for the two gesture orders. A comparison between the simulated means for change in overall entropy and change in MI for both head nouns and dependents can be seen in Figures 3.G.1-3.G.6. As the z-score calculations for the experiment confirmed, all experimental values are significantly different from the simulated means except the change in overall entropy for the majority postnominal condition.

![Figure 3.G.1: Distribution based on simulated means of change in overall entropy across conditions, including visual comparison between simulated mean of means and observed experimental mean for the experiment in Chapter 3. Experimental mean is significantly different from simulated mean of means.](image)

### 3.H Markov chain analysis

In Chapter 1 I discussed experimental and computational work which aims to simulate cultural transmission of linguistic systems and structure. Many of these studies employed a methodology called iterated learning, a process in which the behavioural output of an individual (or several) in generation $n$ is used to construct the training/input for individuals in generation $n+1$. This methodology serves as a useful proxy for examining how behaviours propagate in a population.
over time. An interesting feature of many of these studies is that they often approximate cultural transmission as chains with only one individual in each generation, or at each time-step. As a result, this conceptualisation of iterated learning is equivalent to a finite Markov process (Griffiths & Kalish, 2007) allowing for simple simulations of possible end-states of a process of cultural transmission.

A Markov process, or Markov chain, is designated as a discrete-time stochastic process over a sequence of possible states, or values, of a variable. Crucially, this process is “memoryless”, such that the probability of the random variable \( v_t \) adopting a specific value \( (v_{t=1}, v_{t=2}, ..., v_{t=n}) \) depends only on the present state, and not any prior states (Chung & Walsh, 2006; Norris, 1998). In this way, if a learner is defined by a Markov process, they only have access to the behaviour or knowledge of one individual from the previous generation \( (n-1) \), and has no knowledge of the states of the generation prior to that \( (n-2) \). This type of Markov process can thus be formalised in the following way:

\[
P(v_{t=1}, v_{t=2}, ... v_t = n) = P(v_t | v_{t-1})
\]

(Recreated from Papoulis (1965, p.535))
Figure 3.G.3: Distribution of simulated means of change in MI based on head noun across conditions, including visual comparison between simulated mean of means and observed experimental mean for the experiment in Chapter 3. Experimental mean is significantly different from simulated mean of means.

The possible values of $v$ (the random variable) define the state space for the Markov process, and the probabilities with which one value of $v$ leads to every other available value of $v$ constitutes a transitional matrix between states (Norris, 1998, p.1-3). Generating a stationary distribution based on a Markov process acts as a proxy for a typological distribution over cultural variants (i.e. states) over generations of learners. In this instance I computed the stationary distribution of the experimental transition matrix using eigen decomposition, as this matrix fulfils the criteria for being ergodic, i.e. it is irreducible (every state can be reached by every other state) and aperiodic (Griffiths & Kalish, 2007).

This Markov chain allows us to simulate the iterated learning model of cultural evolution where the data from one generation becomes the training input to the next generation. The parallels between this model and the Markov chain lies in the fact that the sequence of variants represents the chain of successive generations, and the state at the $nth - 1$ variant is what influences the $nth + 1$ variant.

To reiterate, a Markov process is specified by a transition matrix, which defines the probabilities of each possible input state ($S_i$) to transition to each possible output state ($S_j$) after one time step ($t$). The input states in our model were thus defined as the experimental conditions in the experiment in Chapter 3 such that $S_0, S_1, S_2, S_3 = \{\text{natural, unnatural, majority prenominal, majority postnominal}\}$.
where $S_{t-1}$ is the input state and $S_t$ is the output state. Each state has the same majority gesture order for each dependent type as the corresponding experimental condition. Participants were classified as belonging to a specific output state based on which input state that their testing phase data most closely followed in terms of majority order for each dependent type. As such, a participant who produced majority prenominal order across both dependent types would be classified as belonging to the majority prenominal ($S_3$) output state, whereas a participant who produced majority prenominal orders for genitive trials and majority post-nominal orders for adjectives would be classified as belonging to the natural ($S_1$) output state. Data from participants who did not have a clear majority order for either dependent type could not be classified as belonging to any state and so data from these participants were excluded ($N=42$ out of 200). This classification acted as the basis for estimating the probabilities of the transition matrix at the initialising stage (i.e. time 0). Figure 3.H.1 shows the probabilities of the transition matrix used to initialise the finite state Markov chain.

**Stationary distribution**

While the transition matrix shows the probability distribution on the states at time 0, at each time step ($t+1$) the distribution over states evolves. This change
Figure 3.G.5: Distribution of simulated means of change in MI based on dependent type across conditions, including visual comparison between simulated mean of means and observed experimental mean for the experiment in Chapter 3. Experimental mean is significantly different from simulated mean of means.

means that the probability of transitions between some states change to become more or less likely than at $t - 1$, and this change is determined by $P$. The stationary distribution of a Markov chain describes the distribution over states $(X_t)$ after a sufficiently long time that the distribution of $X_t$ no longer changes. Thus if we take $\pi$ to be a column vector of probabilities over possible states that a Markov chain can take. Then, $\pi$ is the stationary distribution if it has the property:

$$\pi^T = \pi^T P$$

As such the vector $\pi$ is a left eigenvector of $P$ with eigenvalue equal to 1 and so I computed the stationary distribution using the *eigenvector* functions in base R. The stationary distribution in figure 3.H.2 thus shows the expected proportion of each output state (i.e. language type) after sufficient time has passed for the distribution to stabilise. This distribution shows that the highest number of languages are expected to be those where both adjectives and genitives precede the head noun, followed by those where adjectives precede and genitives follow. The least common languages are expected to be those which are natural (prenominal genitives and postnominal adjectives) and those where both dependents are postnominal. This is in direct contrast with what we see in the typology as the most
common language patterns for these dependents are harmonic postnominal and natural, followed harmonic prenominal and finally unnatural languages.

The validity of these results are difficult to evaluate as there was such a high proportion of data loss, going from 200 participants in the original experiment to 158 participants who could be classified for this analysis (loss = 21%). This data loss was fairly similar across conditions, ranging between 9 to 12 participants lost per condition (see Table 3.H.1). This is clearly a drawback of the classification system used to form the basis of the input states, yet no other suitable alternative was available due to the nature of the task. In addition, the analysis only estimates the probability of four states, despite the fact that the states can vary widely beyond those four states. The proportion of orders for the two dependent types do not just take on the value “majority” and “minority” in participants productions. As a result, the reliability of the stationary distribution being a reflection of the pressures involved in the experimental task is questionable, hence why this analysis was not repeated for any of the other regularisation experiments in this thesis.
Figure 3.H.1: Representation of initialising transition matrix underlying Markov chain process based on testing data from the experiment in Chapter 3. Input states (columns) represent conditions participants were trained in, output states (rows) represent classifications of their output. Probabilities are derived based on how likely participants in any given input condition are to transition to each output state. Darker colours show higher probabilities, the diagonal line of slightly darker squares show that participants were most likely to remain in the same state between input and output.

Figure 3.H.2: Stationary distribution across the four language types, extracted based on transitional matrix from the experiment in Chapter 3 using eigenvector functions. The two most common language types are prenominal harmonic languages and unnatural languages.
Table 3.H.1: Number of participants per condition in the experiment in Chapter 3 and number participants that could be classified for Markov chain analysis

<table>
<thead>
<tr>
<th>Condition</th>
<th>Experiment</th>
<th>Classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>50</td>
<td>38</td>
</tr>
<tr>
<td>Unnatural</td>
<td>53</td>
<td>42</td>
</tr>
<tr>
<td>Majority prenominal</td>
<td>46</td>
<td>37</td>
</tr>
<tr>
<td>Majority postnominal</td>
<td>51</td>
<td>41</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>158</td>
</tr>
</tbody>
</table>
Chapter 4

Does degree of innovation modulate the influence of cognitive biases on language structure?

4.1 Preface: Chapter 4

This last main content chapter of my thesis was designed to explore the more central regions of the Scale of Innovation. I did this because the way that the category-specific preferences for postnominal adjectives and prenominal adjectives influence typological structure still remained somewhat mysterious after the initial three experiments. Without clear behavioural effects during learning, the impact of category-specific biases which may have been active during language creation should have been washed out by the iterative effect of system-wide biases like harmony. The fact that they are still readily observable in typology of both spoken and (to some extent) signed languages thus remained unexplained. However, as discussed in Sections 1.5.2 and 1.5.3 of Chapter 1, other contexts and language tasks combine the need to both learn parts of a language system and innovate new parts of that language system. These combined tasks reside in the middle of the Scale of Innovation and are much more common in natural language situations than either pure learning/memorisation or pure improvisation contexts as we continuously acquire new words or combine words in novel structures.

With this in mind, I designed a set of two experiments that would let me compare the effect of category-specific biases across two artificial language learning tasks, one regularisation experiment where participants were trained and tested on the same items, and one where the testing items were novel. The second experiment thus required generalisation at test, a context which involves higher amounts
of innovation (i.e. interacting with novel linguistic material). Experiment 1 is similar to the one in Chapter 3, but with a more limited set of meanings across the two dependent types. This restricted set of meanings allowed me to perfectly match the training phase for both experiments in this chapter, making the only difference between Experiment 1 and 2 be the items they are presented with during the testing phase. The first experiment therefore partly acts as a replication of the experiment in Chapter 3, and as the basis of comparison for the second experiment in this chapter. Differences in learning behaviour and the influence of naturalness can thus be compared more readily between these two experiments.

The results of Experiment 1 replicate the main findings from Chapter 3, finding no reliable learning advantage for orders favoured by the category-specific biases. There were however consistent effects of both harmony and several strategies for reducing unconditioned variation, supporting the notion that system-wide biases influenced participants’ behaviour at test. Experiment 2 showed that participants chose more orders favoured by the category-specific biases than is expected by chance, however when comparing the learning behaviour of participants across the two experiments there was no evidence that natural orders in Experiment 2 were learned better than natural orders in Experiment 1. In fact, when data from both experiments were included natural orders were learned better than unnatural ones across the experiments.

Once again, the appendix for this chapter includes details about the Monte Carlo simulations used to validate the reliability of the changes in entropy and Mutual Information that were observed in this chapter. Furthermore, the appendix also provides a summary of the additional spoken languages that participants reported having any knowledge of across both experiments.
4.2 Holtz et al. (under review): author contributions

The following paper is an expanded version of an article which was submitted to Language in February 2024. The paper was co-authored with my two supervisors, Jennifer Culbertson and Simon Kirby. The development of the hypotheses, experimental design, and analysis took place during supervision meetings where all three authors were present and contributing. Both co-authors gave feedback during the write-up process. The manuscript has been reformatted to adhere to thesis formatting guidelines. Discrepancies in general content between this paper and any future published version may occur as a result of the review process.
4.3 Abstract

A number of previous studies have argued that individual-level cognitive biases can influence population-level typological patterns. Here, we highlight a recurring pattern among these studies: different types of cognitive biases appear to be most active in particular tasks or contexts. For example, biases which affect the organisation of linguistic items into a wider system are especially active in language learning tasks; biases which influence the expression of individual linguistic items or categories independent of any system tend to be observed in language creation tasks. The contexts in which a cognitive bias applies are important: the more restricted the contexts, the greater the limit placed on that bias to influence language structure. Item-based (or category-specific) biases may only influence behaviour when a language is first created, or when novel items are needed. By contrast, system-wide biases can apply iteratively with each new generation of learners. This study seeks to further examine the apparent contextual restrictions on category-specific biases. We target linguistic tasks that may combine elements of language learning with language innovation—for example contexts which require generalisation to new items. Our goal is to see whether both types of biases emerge in this context. If so, this might supply a missing link, explaining how category-specific biases come to influence typology. Using word order in the noun phrase as a case study, we examine whether category-specific biases favouring postnominal adjectives but prenominal genitives—previously only seen in improvisation tasks—also influence behaviour in a task that involves generalisation.

4.4 Introduction

A variety of cognitive biases have been proposed to influence the way people interact with linguistic stimuli. These biases are argued to be active during language acquisition (Chater & Vitányi, 2003; Christiansen, 2000; Culbertson et al., 2012), production (MacDonald, 2013), communication (Fedzechkina et al., 2023; Ferdinand et al., 2019), processing (Gibson et al., 2019; Hawkins, 1983), and the creation of new linguistic systems (Goldin-Meadow et al., 2008; Schouwstra et al., 2020; Senghas et al., 2014). Here, we are interested in two of these contexts in particular: biases active during learning, and biases active during language creation. As we outline below, there is evidence that these biases target distinct types of linguistic phenomena—patterns that hold across a language system, and those that hold of a specific type of item or category of items—and that these
biases might differ in their ability to shape language typology. While the former can apply iteratively, at each new generation of learners, intuitively the latter context appears more restricted. Whether item or category-specific biases are indeed restricted to contexts of language creation, or whether these biases might emerge in a wider-range of contexts is the main question we address here. Before describing our approach, we first summarise the special role of learning biases in shaping typology, and the evidence to-date that system-wide and category-specific biases are distinct.

Biases active during language learning have received special attention, due to the fact that they represent a potential probabilistic solution to the problem of induction that children face during language acquisition (Culbertson, 2012). This problem relates to the fact that the space of hypothetically possible language grammars is very large, and yet children are able to identify and internalise the grammar that underpins their native language(s) based on linguistic input which can never fully rule out all other available hypothetical grammars (Berwick et al., 2011; Chomsky, 1965; Pinker, 1987; Wolff, 1982). This problem, sometimes referred to as the “poverty of the stimulus”, is one of the main motivations behind the emergence of generative grammar. The generative approach to language acquisition argues that this problem is solved by humans possessing an innate language faculty which places inviolable constraints on the shape of possible languages. During acquisition, the language faculty thus rules out any hypothetically possible grammars which violate these constraints (Chomsky, 1965, 1966).

Early typological work which identified language universals (absolute constraints on cross-linguistic structure) seemed to support this view since some features, or combinations of features appeared to be categorically absent from typological samples (Greenberg, 1963; Hawkins, 1983). Much work in the generative tradition has continued to search for such cases (e.g., Baker, 2001; Biberauer et al., 2014; Cinque, 2005; Prince & Smolensky, 2004, among many others). The identification of such universals is expected if the structure of human grammars are determined by inviolable constraints. However, many proposed absolute universals have since been found to have exceptions (for a summary of some of these exceptions, see Evans and Levinson (2009)).

These discoveries have prompted alternative views aimed at explaining, not the impossibility of particular linguistic patterns, but instead differences in likelihood (Christiansen & Chater, 2016; Dryer, 1992; Evans & Levinson, 2009; Gibson et al., 2019; Hahn et al., 2020; Kirby et al., 2015; Thompson et al., 2016). These alternatives have in common the view that the system which gives rise to statistical typological tendencies is expected to itself be probabilistic in nature. For
example, rather than hard-and-fast constraints on the linguistic system, some of these accounts posit cognitive biases which can impact cross-linguistic structure during language acquisition (Culbertson, 2012). For example, conceptualised as parallel to Bayesian priors, these cognitive biases increase the prior probability of certain hypotheses (i.e., certain grammars) being entertained by learners during acquisition (Culbertson & Kirby, 2016; Culbertson et al., 2013; Perfors et al., 2011). Cognitive biases make bias-congruent grammatical systems more likely to be acquired by learners, thereby creating the basis for the existence of statistical typological patterns. Importantly, cognitive biases active during language learning hold a special status as typology-shaping biases since these biases have the opportunity to affect linguistic behaviour with each new generation of learners, including children acquiring their first language and adults acquiring a second language.

Examples of biases argued to be active during learning include a bias for word order harmony (i.e., consistent order between syntactic heads and dependents), and a bias for phonological alternations/rules that involve fewer feature changes. Evidence for a harmony bias comes from, for example, artificial language learning tasks where participants tend to learn harmonic languages (either head-initial or head-final) better than non-harmonic ones (Culbertson, Franck, et al., 2020; Culbertson & Newport, 2015; Culbertson et al., 2012). Similarly, evidence for a preference for phonological alternations implicating fewer features also comes from artificial language learning experiments (see Moreton & Pater, 2012, for a review). These experiments have consistently found evidence that rules which require the encoding of fewer phonological features are easier to learn (Conrad et al., 1977; Pycha et al., 2003; Saffran & Thiessen, 2003; Skoruppa & Peperkamp, 2011).1 Crucially, both the bias for harmony and the bias for phonological rules implicating fewer features can be conceived of as system-wide organisational principles. Both suggest that rules should apply more widely or generally, across a system. Indeed, these two system-wide biases have been argued to be attributable to the same basic cognitive principle, namely the human propensity to default to the simplest possible explanation for the available data (Chater & Vitányi, 2003). A simplicity bias active during learning has wide-ranging implications

\[\text{For example, an alternation which targets all high vowels (+high) is easier to learn than a rule that targets all high rounded vowels (+high and +rounded). Similarly, rules which require permutations of fewer features are easier to learn than rules requiring the permutation of several features (Peperkamp et al., 2006; Skoruppa et al., 2011; White, 2014). For example, changing one feature value is easier than changing several, meaning that learning a rule which changes [t] to [d] (change in feature [voice]) is easier than learning a rule which changes [t] to [b] (change in feature [voice] and [continuant]).}\]
for language across many domains, and features prominent in explanations of
many typological tendencies (e.g., Culbertson & Kirby, 2016; Kirby et al., 2015;

Importantly, in contrast with biases active in learning, biases affecting be-
behaviour during the creation of linguistic systems appear to be more restricted;
they emerge in tasks involving improvisation or language creation rather than
learning. This suggests they may have only limited scope to shape language
structure (Schouwstra et al., 2020). For example, there is evidence that the basic
word order patterns Subject-Object-Verb and Subject-Verb-Object are typolog-
ically more prevalent than alternatives, and preferred by participants in exper-
imental tasks involving improvised gesture (Goldin-Meadow et al., 2008; Hall
et al., 2014; Motamedi, Wolters, Naegeli, et al., 2022; Schouwstra & de Swart,
2014). Several studies have also examined patterns of word order in the noun
phrase. For example, across languages, adjectives tend to appear linearly closest
to nouns, followed by numerals, with demonstratives furthest away (e.g., ‘these
two blue vases’ or ‘vases blue two these’). These same orders have been shown
to be preferred in experimental tasks where participants learn only the order of
nominal dependents relative to the noun, but must infer (or guess) which are lin-
eary closest (Culbertson, Schouwstra, & Kirby, 2020; Martin et al., 2019, 2020,
2024). There are also distinct tendencies among different dependent types to
be either prenominal or postnominal. For example, adjectives tend to be post-
nominal, but genitives tend to be prenominal. Participants tend to prefer these
orders in improvisation tasks, where participants create new linguistic expressions
to signify these meanings (e.g., Culbertson, Schouwstra, & Kirby, 2020; Jaffan
et al., 2020) and in gesture judgement tasks, where gesture sequences conveying
images are presented in isolation (e.g., Holtz, Kirby, & Culbertson, 2023). These
biases, found in tasks involving some kind of improvisation, are not about the
organisation of items into a grammatical system. Rather, they are about partic-
ular categories in a single utterance. Unlike the system-wide bias for harmony,
for example, for these biases to apply, nothing about how the rest of the system
works is necessary. To determine whether a given phrase conforms to a preferred
order (e.g., Subject-Object-Verb, or Noun-Adjective-Numeral-Demonstrative, or
Genitive-Noun ), only this phrase is relevant. These are item- or category-specific
biases.

The studies described above thus target different types of biases— system-
wide biases versus category or item-specific biases—and they use different tasks—
learning versus improvisation. Is this an accident? In fact, there is some evidence
that specific biases are restricted to specific task types. For example, in impro-
193
visation tasks where participants can freely place noun phrase elements there is no evidence for an additional preference for harmony (Culbertson, Schouwstra, & Kirby, 2020). This suggests that, while the system-wide preference for harmony is observed in learning, it is not necessarily observed in pure improvisation. Similarly, while category-specific preference for nominal dependent ordering is evident in improvisation tasks, consistent evidence that these same orders are learned better than others is missing (Holtz, 2019; Holtz, Culbertson, & Kirby, 2023). In this paper, we argue that this pattern of results, where some biases seem to be active during language learning, and some biases seem to be active during language creation, is potentially both more significant and more general than has previously been noted.

4.4.1 Scale of Innovation

The two types of tasks, or contexts, which we have discussed so far, language learning and language creation/improvisation, can be characterised as opposite ends of a scale, or spectrum, based on the amount of innovation that is required in each type of task. We introduce this spectrum in Figure 4.4.1 as the Scale of Innovation. On this scale, contexts that require mainly learning, a little or no innovation, are on the far left. For example, imagine a conventionalised language system, fully established prior to learning, for which evidence is readily available. In this (idealised) context, the system, or parts of it, can in principle be memorised wholesale. This might roughly correspond to an early stage of natural language acquisition, where item-based memorisation (e.g., of multi-word phrases or chunks) initially dominates, and representations which allow generalisation (i.e., of rules of compositional structure) only come later (as argued by, e.g., Arnon, 2021; Lieven et al., 2003; Tomasello, 2000). Wholesale memorisation of an established system is also possible in the context of experimental tasks involving learning of miniature artificial languages, where the entire system can be presented to participants (e.g., Christiansen, 2000; Tabullo et al., 2012; Tily et al., 2011).

Next along the scale are what we are calling regularisation contexts, where the presence of irregularity and/or unconditioned variation means that the system is less firmly established. In this context learners “discover” regularities in the system, and therefore introduces an element of innovation: learners can introduce even more regularity into the system. Conditioned variation is very common in natural language, and unconditioned variation may occur during periods of language change, where different variants arise and compete with one another. Thus
learners are likely to be faced with this type of context. Similarly, these contexts correspond to experiments which introduce linguistic variation in order to observe how participants might change the distribution of variants in the system (e.g., Culbertson et al., 2012; Hudson Kam & Newport, 2005). Notably, in these regularisation contexts, the variants have already been established, and thus the opportunity for innovation of new expressions remains restricted.

Figure 4.4.1: Scale of Innovation with contexts of language learning and use requiring increasing amounts of innovation as you move rightward along the scale. Length of lines show how much of the language system that individuals in each type of context have access to (e.g. a full line across the scale means the whole system is available).

In the middle of the scale are contexts that involve generalisation, in other words the application or extension of established patterns in the system to new linguistic items and situations. These contexts require some innovation, since there is not necessarily direct evidence for how new items should be integrated into the linguistic system, only indirect evidence from other previously learned items (e.g., of the same type). These contexts are arguably pervasive during natural language learning and use, whenever we encounter novel types of known categories. Generalising tasks have also been used in experimental paradigms as measure of more robust or abstract learning beyond memorisation (e.g., Martin & Peperkamp, 2020; Pycha et al., 2003).

Higher still on the scale are contexts where people must extrapolate. In these contexts, the evidence present in the input for potential pattern extension, but in relation to new items or structures is less clear. For example, as adults we learn new words and phrases that need to be integrated into our existing language system by analogy with existing rules, categories, or items. Some of these might be quite similar, and thus generalisation is required. But others might be quite distinct (e.g., in cases of loan-word adaptation). Similarly, children do not learn all linguistic categories or constructions simultaneously (Gentner, 1982; Tardif et al., 1999), but must integrate these into an existing system which already might
include similar constructions and categories or not. Extrapolation is often used in experimental tasks as well, in order to see how participants use limited or ambiguous evidence to generate novel items/structures or treat novel categories (e.g., Culbertson & Adger, 2014; Finley, 2018; Saldana et al., 2021; Wang et al., 2023).

Finally, at the far right of the scale are improvising contexts. These include instances of novel language creation, where linguistic expressions need to be innovated in real-time, either without any pre-existing conventions at all, or in cases where the relationship between the novel item or expression and any existing system or conventions is highly under-determined. Although neologisms are certainly created in established languages, these situations are arguably much rarer in natural language learning and use compared with the others on the scale. Improvisation may have been very common during the early stages of language evolution in humans, and the closest modern equivalent of innovation without any conventionalised system of communication in place is the emergence of new sign languages in deaf communities that previously had no shared language, such as the recent emergence of Nicaraguan Sign Language (Senghas et al., 2014). As discussed above, these contexts are often used in experimental paradigms using gesture-based improvisation (e.g., Goldin-Meadow et al., 2008; Hall et al., 2014; Motamedi, Wolters, Naegeli, et al., 2022; Schouwstra & de Swart, 2014).²

4.4.2 Category-specific biases across the scale

We can now return to the main question of interest in this paper: the contexts in which category-specific biases emerge. As discussed in Section 4.4, these biases appear to align with typology (at least in some cases), and yet most experimental evidence for them comes only from tasks on the higher end of the Scale of Innovation, but not the lower end. While there is evidence that weak biases active during learning can be amplified by cultural transmission (Reali & Griffiths, 2009; Thompson et al., 2016)—i.e., since the bias can affect language learning again at each generation—this is less clear when the bias affects improvisation. If we mainly find evidence for these category-specific biases in contexts at the far right of the scale, then it is surprising that we nevertheless see evidence for

²Gesture is not the only modality which has been used in pure improvisation experiments. For example, there are studies in which participants innovate new pictorial (Garrod et al., 2010) or non-linguistic auditory (Fay et al., 2013, 2014; Perlman et al., 2015) expressions to signify different meanings. It is worth noting that there are also many learning-based studies using novel communication systems which are very different from human languages, including whistles (Verhoeff et al., 2014) or layers of images that together make up a scene (Goldin-Meadow et al., 2008; Struhl et al., 2017).
them reflected in the typology. If, on the other hand, these biases are at play in other contexts, perhaps those requiring some level of innovation, then it is much less mysterious how they come to shape typology. In fact, there is already some evidence suggesting that both system-wide biases and category-specific biases can “bleed” into the middle regions of the scale under specific conditions. This makes sense, since the contexts discussed above are conceptualised as existing on a scale precisely because the boundaries between regions of the scale are fuzzy. Both natural language learning and experimental tasks often involve a combination of these.

For example, Wang et al. (2023) found that participants will extrapolate the word order from one phrase type to another as long as the meanings involved were sufficiently similar. This suggests that a system-wide bias like harmony applies not just in learning, but also in contexts that are higher on the Scale of Innovation. At the same time, there is some evidence for category-specific biases emerging in tasks that require both learning and generalisation. Do et al. (2022) use a regularisation design but also include novel items in the test set. They find some evidence of a category-specific preference for postnominal adjectives, unlike Holtz, Culbertson, and Kirby (2023), who do not use novel items at test.

The goal of this study is to probe further the possibility that category-specific biases are at work in the centre of the Scale of Innovation. We use as a test case word order in different nominal dependents, specifically among adjectives and genitives. This case is particularly interesting, since both system-wide and category-specific biases are at play, and pull language structure in opposite directions. As shown in Table 4.4.1, adjectives exhibit a strong postnominal tendency, and genitives show a clear (but less strong) prenominal tendency, across both spoken and signed languages. Furthermore, as Table 4.4.2 shows, when both dependent types are taken into account, the order which follows these category-specific biases is very common. However, so are both harmonic patterns, where one dependent-type does not follow its category-specific preference. Building on previous experimental work which has focused on tasks closer to the end points of the Scale of Innovation, and has identified both types of biases at either end (Holtz, Culbertson, & Kirby, 2023; Holtz, Kirby, & Culbertson, 2023), we test what happens when generalisation is required. In Experiment 1, participants are trained and tested on the same linguistic item, whereas in Experiment 2 they are trained and tested on different items. This allows us to compare how these two types of biases are at work as the experiment context moves along the Scale of

\[3\] For example, they found that participants extrapolate the order of verbs to adpositions and adjectives with more verb-like meanings, but not to colour or texture adjectives.
Innovation.

Table 4.4.1: Typological counts of the order of nouns with adjective and noun with genitive, in spoken (spo.) and signed (sig.) languages (Coons, 2022; Dryer, 2013a, 2013b).

<table>
<thead>
<tr>
<th>Adjective</th>
<th>N (spo.)</th>
<th>N (sig.)</th>
<th>Genitive</th>
<th>N (spo.)</th>
<th>N (sig.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postnominal</td>
<td>879</td>
<td>16</td>
<td>Prenominal</td>
<td>468</td>
<td>2</td>
</tr>
<tr>
<td>Prenominal</td>
<td>373</td>
<td>10</td>
<td>Prenominal</td>
<td>685</td>
<td>22</td>
</tr>
<tr>
<td>Other</td>
<td>110</td>
<td>15</td>
<td>Other</td>
<td>96</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 4.4.2: Typological counts of the order of noun, genitive and adjective, based on spoken (spo.) language typological data (Dryer, 2013a, 2013b) and signed (sig.) language typological data (Coons, 2022). Harmonic patterns are common, but so is a particular non-harmonic ordering, with postnominal adjectives and prenominal genitives (bolded).

<table>
<thead>
<tr>
<th>Order</th>
<th>Noun-Adjective</th>
<th>Adjective-Noun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modality</td>
<td>spo.</td>
<td>sig.</td>
</tr>
<tr>
<td>Noun-Genitive</td>
<td>342</td>
<td>-</td>
</tr>
<tr>
<td>Genitive-Noun</td>
<td>342</td>
<td>9</td>
</tr>
</tbody>
</table>

4.5  Experiment 1

In Experiment 1 we first aim to replicate Holtz, Culbertson, and Kirby (2023), who find evidence for a system-wide bias for harmony, but no influence of category-specific biases (either postnominal adjectives or prenominal genitives) in a learning task. We will then use the same design to move further along the Scale of Innovation to a task with generalisation and compare the results across the two studies.

4.5.1  Method

We use an artificial gesture learning task in which participants are trained on a word order system with variation. The task is therefore a regularisation design (low-innovation context). In the testing phase participants must chose between two gestural expressions of a specific meaning. This basic methodology was inspired by previous work examining the way that some category-specific cognitive biases (for basic word order) influence learning behaviour (Motamedi, Wolters, Schouwstra, & Kirby, 2022). Regularisation tasks, as mentioned above, purposefully teach participants a system with some variability in it, and compare how
participants change or recreate that variability in the system they themselves construct during testing (Culbertson, 2023; Culbertson et al., 2012). We measure the degree to which participants maintain or change the variability that they were exposed to.

Participants were exposed to a majority and a minority gesture order for both adjective and genitive meanings during training. 75% of the time participants saw the majority order, and 25% of the time they saw the minority order. There were four different training conditions, which manipulate whether the majority order is prenominal or postnominal for a given dependent type. Table 4.5.1 shows the distribution of orders that participants were exposed to during training in all four conditions. In two of the conditions, the training data was majority prenominal or postnominal across both dependent types. In the natural condition the majority order for adjectives was postnominal, and the majority order for genitives was prenominal (following the category-specific biases). The unnatural condition was the opposite of this.

Table 4.5.1: Percentage of prenominal and postnominal gesture orders for each dependent type in the training data across the four experimental conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Adjective</th>
<th>Genitive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Post</td>
</tr>
<tr>
<td>Natural</td>
<td>25%</td>
<td>75%</td>
</tr>
<tr>
<td>Unnatural</td>
<td>75%</td>
<td>25%</td>
</tr>
<tr>
<td>Majority prenominal</td>
<td>75%</td>
<td>25%</td>
</tr>
<tr>
<td>Majority postnominal</td>
<td>25%</td>
<td>75%</td>
</tr>
</tbody>
</table>

Materials

The experiment was programmed using the jsPsych library (de Leeuw et al., 2023) to run in participants’ web browser. Materials for the study included a set of grayscale digital drawings which were used to signify possessive (genitive) and descriptive (adjective) meanings. The genitive images showed two different possessors, a cyclops and a vampire, owning items such as a hat, a scarf, a book, and a cup. The adjective images showed those same items but with two different patterns, being either spotted or striped. The full combinations of these adjectives, genitives and nouns make a total set of 16 meanings, depicted in 16 images.4

Each meaning was signified using a pair of gesture videos, each consisting of two gestures. In one video the first gesture signified the head noun, and the

4All stimuli, experiment code, data and analysis scripts can be found at this link: https://osf.io/dep4s/?view_only=bd48772f485a11243207b4e6e6e6e
second gesture signified the dependent (postnominal order). In the other video the first gesture signified the dependent, and the second gesture signified the head noun (prenominal order). These gestures in the videos were performed by a model gesturer and in total they performed two gesture sequences for all 16 meanings, making a total of 32 stimuli videos. All gestures were two-handed and performed to the beat of a metronome to match the beginning and end of each gesture. Each gesture video started and finished with the model gesturer’s hands in a neutral position, clasped in front of them, and were exactly 4.389 seconds long.

**Procedure**

The initial instructions included the consent form and information sheet for the study, where participants were told that the study involved them learning how to describe objects and how to express ownership in a made-up sign language. After participants had provided consent they were then familiarised with the type of stimuli they would see throughout the study. Participants were randomly assigned to one of the four conditions in Table 4.5.1, and a pseudo-randomised stimuli set consisting of one “worn” item (*hat* or *scarf*) and one “held” item (*cup* or *book*), combined with one adjective (*striped* or *spotted*) and one genitive (*cyclops* and *vampire*). There were two familiarisation trials, these displayed a 2x2 grid of images, one showing four genitive meanings and the other showing four adjective meanings that contrasted along both the noun and dependent dimension. Figure 4.5.1 shows two example grids from these trials.

The next phase was the training phase. Participants were told that they would see the same kinds of image grids, but with a target meaning highlighted, and that videos showing someone expressing the target meaning would play under the images. They were told to “sit back and watch carefully.” The training trials started with the same kind of 2x2 image grid being displayed but with one of the images highlighted. The highlighted image was the target image, and below them a gesture video appeared half a second later and played all the way through. The trials progressed automatically and there were a total of 32 training trials (eight per target meaning in the participant’s stimuli set). The proportion of gesture orders that participants saw varied by condition, but the majority order for a given meaning was always displayed six times and the minority order twice.

The testing phase included the same kinds of grids with a target highlighted as in the training conditions, but this time participants were shown two videos under the grid. These videos were always the prenominal and postnominal gesture
orders for the target meanings and they were horizontally aligned and the left-right position randomised per trial per participant (see Figure 4.5.2 for an example of a test trial). The videos looped simultaneously and participants were told “In this second part of the experiment, you will be shown the same collection of images again as well as two gesture videos. Please click on the corresponding gesture video like you saw in the first part.” The videos looped until a selection was made. Between each trial participants also clicked a centred “Next” button to reset the mouse and progress to the next test trial. The number of test trials matched the training trials, giving a total of 32 data points, with eight selections per target meaning.

Finally, participants answered two brief demographic questions. The first asked if they were proficient in a sign language, and (if they were) which one. The last question asked them to list all the spoken languages they know and estimate their proficiency in these languages on a scale of 1-10 (with 10 being the highest proficiency).

**Participants**

213 participants were recruited using the online crowdsourcing platform Prolific. We employed preregistered prescreening criteria which limited the participant pool to those who reported English as their first language, who had 95% previous task-approval rate, and who had not participated in any of our previous experiments or pilots.\(^5\) The study took approximately ten minutes to complete and the payment rate was set to the equivalent of £10.42 per hour based on an

---

\(^5\)Preregistration for this project can be found at this link: https://osf.io/mq8s/?view_only=0719fcf208eb4a4ebbb240cb61df7d4af0
Figure 4.5.2: Sample trial from the testing phase for the target meaning *spotted hat*. The two video choices include the prenominal and postnominal gesture orders for this meaning.

estimated completion time of ten minutes. Following preregistered exclusion criteria we excluded seven participants who reported proficiency in at least one sign language, as well as two participants who gave incoherent responses to the survey questions. Finally, a further five participants’ data was lost due to connection issues during study completion. The data in the final analysis comes from 50 participants in the natural condition, 51 in the unnatural condition, 50 in the majority prenominal condition, and 49 in the majority postnominal condition.

### 4.5.2 Results

Based on Holtz, Culbertson, and Kirby (2023), we had three main predictions for this experiment. Firstly, we predicted an overall preference for harmony, such that participants across conditions would prefer consistent orders between heads and dependents. We also expected to find evidence of other variability-reducing strategies, such as conditioning order on specific lexical items. Importantly, based on the results from Holtz, Culbertson, and Kirby (2023) and the conceptualisation of the Scale of Innovation, we do not expect to find any evidence of a preference for natural orders. This means we do not expect better overall learning or more

---

6 Incoherent here meant answers that included no discernible words, e.g. participants who pressed letters randomly.
regularisation in the natural condition, or better learning of individual dependent orders that conform to a category-specific bias.

Learning

An overview of the output systems generated by participants in all conditions in Experiment 1 can be seen in Figure 4.5.3. We first analyse the degree to which participants accurately learned the majority orders they were trained on—i.e., whether the orders they chose in the testing phase matched the orders they saw most in training. We ran a mixed effects logistic regression model with the binary outcome variable *majority order* (1 indicating a match between participants’ selected order and the majority order in the training phase, 0 indicating a mismatch). The fixed effects were *dependent type* (either adjective or genitive) and *condition* (natural, unnatural, majority prenominal, or majority postnominal), both of which were deviation coded. The random effect structure included a by-participant random intercept, and a random slope for dependent type. The model had a significant positive intercept ($\beta = 1.21, SE = 0.12, z = 9.82, p < 0.001$), suggesting that participants across all conditions, on average, chose the majority order they were trained on at a rate above chance. This trend is also visually present in Figure 4.5.4, which plots proportion of majority orders chosen by participants during the testing phase, split by dependent type. Model comparison using a likelihood ratio test found that the intercept only model was the best fit for the data, and that including condition, dependent type or their interaction did not improve model fit ($\chi^2 = 2.60, p = 0.46; \chi^2 = 0.41, p = 0.52; \chi^2 = 0.89, p = 0.83$). This suggests that there was no significant difference in learning behaviour across conditions, nor were there differences in learning the order for adjectives or genitives.

Naturalness

The learning data already suggest that naturalness does not have an effect in this study. However, we can also specifically examine participants’ tendency to select natural orders (postnominal order for adjectives and prenominal order for genitives). Figure 4.5.5 shows the individual and overall proportions of natural orders that participants selected, split by dependent type. We examined the overall tendency for choosing natural orders, as well as differences between conditions using a mixed effects logistic regression model with *natural order* as the binary outcome variable (1 indicating a match between participants’ selected order and the natural order for that dependent type, 0 indicating a mismatch). Recall that
Figure 4.5.3: Proportion of postnominal orders selected by each participant for both dependent types across all conditions in Experiment 1. Each small shape represents a participant, and the four larger shapes represent the input proportions in each condition.

The natural order is postnominal for adjectives and prenominal for genitives. We included the same fixed effects and random effects as in the learning model above. As the overall mean in Figure 4.5.5 indicates, the intercept was non-significant, meaning that participants across the experiment did not select natural orders at a rate above chance. However, model comparison with a likelihood ratio test found that the model including an interaction between condition and dependent type improved model fit ($\chi^2 = 53.74, p < 0.001$). There was a positive effect of condition, such that participants in the natural condition selected more natural orders, compared to the grand mean ($\beta = 1.13, SE = 0.18, z = 6.33, p < 0.001$). Additionally, there was also a negative effect of condition, such that participants in the unnatural condition were less likely to select natural orders, compared to the grand mean ($\beta = -1.02, SE = 0.17, z = -5.85, p < 0.001$). Finally, the interaction between condition and dependent type was significant for both majority conditions. The direction of the interaction effect depended on whether the majority order of the condition matched the natural order for a given dependent. For the majority postnominal condition the effect was negative, as participants chose the natural order for genitives less often, compared to the grand mean ($\beta = -2.67, SE = 0.44, z = -6.04, p < 0.001$). Whereas the effect was positive in the majority prenominal condition, as participants chose the natural order more often, compared to the grand mean ($\beta = 2.85, SE = 0.44, z = 6.45, p < 0.001$).
Figure 4.5.4: Proportion of test trials where participants selected the majority input order for each condition and dependent type in Experiment 1. Right-hand facet shows grand mean. Participants tended to reproduce the majority orders from their training for both dependent types. Error bars show bootstrapped 95% CIs.

These effects reflect the previous learning analysis since participants learned the orders they were trained on they selected more natural orders when the majority order was natural (either for a specific dependent type or across the full condition), and selected less natural orders when the majority order was unnatural (either for a specific dependent type or across the full condition), compared to the grand mean.

Harmony

Recall that in contrast to naturalness, we expect to see a preference for harmony in this experiment. We measured harmony as change in overall entropy between participants’ training data and the gesture orders they selected at test (following other work using the same type of measurements for harmony Ferdinan et al. (2019), Motamed, Wolters, Schouwsstra, and Kirby (2022), and Samara et al. (2017)). A more consistent use of the same gesture order across the full language system would appear as a significant drop in entropy between input and output, i.e. a significant negative value for entropy change. We define the entropy ($H$) or a language system in this study as:
Figure 4.5.5: Proportion of test trials where participants selected the natural order (i.e., prenominal for genitives, postnominal for adjectives) for each condition and dependent type in Experiment 1. Right-hand facet shows the grand mean. There was no overall preference for natural orders. Error bars show bootstrapped 95% CIs.

\[ H(V) = -\sum_{v_i \in V} p(v_i) \log_2 p(v_i) \]

Where \( V \) represents the two possible variants (gesture orders). The input entropy varied between some of the conditions. The natural and unnatural condition had a 50/50 division between prenominal and postnominal orders across the whole training system, meaning that the training data in these condition were maximally entropic for a system with two variants (i.e. had an entropy value of 1). The majority prenominal and majority postnominal conditions both had a 75/25 split across the whole training system, meaning that their starting entropy was slightly lower than the other two conditions at 0.8112781. An increase in harmony (a significant negative entropy change) between training and testing data signifies more use of one of the two gesture orders. Figure 4.5.6 shows the mean change in entropy for each participant, condition, as well as the overall mean change.

We used a combination of tools to evaluate the reliability of these changes.\(^7\) First, we calculated bootstrapped confidence intervals around each condition.

\(^7\)Due to the non-normal distribution of values that our data could take in these experimental conditions we do not report any linear models examining changes in entropy for either of our experiments.
mean using 10,000 samples with the ‘boot’ package in R (Canty & Ripley, 2021). These findings suggested that the mean change was reliable for all conditions. To verify these values we also performed a Monte Carlo simulation, where the probability of sampling prenominal and postnominal order was set to the training proportions for each condition. This allowed us to compare a theoretical baseline of probability matching between input and output to the changes in entropy we observe in the experiment. The simulation ran for 10,000 runs through the experiment, giving us 64 million data points. We extracted mean changes in entropy for each run, as well as mean changes in entropy per condition for each run. Z-scores were then calculated between these simulated means and the means obtained from the experiment. The result of this comparison can be seen in Table 4.5.2 and shows that change in entropy was indeed reliable overall and across all conditions.

![Mean entropy change](image)

**Figure 4.5.6:** Mean change in entropy between input and output, showing participants tend to harmonise across all conditions in Experiment 1. Right-hand facet shows grand mean change in entropy. Error bars show bootstrapped 95% CIs.

**Conditioned variation**

Harmony is not the only available strategy that participants could have used to limit the unpredictable variation that they were exposed to during training. Previous research has found that participants will often introduce lexical conditioning to make variation predictable. In our experiment participants have two options for possible conditioning environments, either the dependent or the head noun.
Table 4.5.2: Experimental means, simulated means and resulting z-scores for change in overall entropy in Experiment 1. Z-scores show that experimental means are reliably different from the simulated means (indicating that entropy dropped significantly) in all conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Exp. mean</th>
<th>Sim. mean</th>
<th>z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>-0.142</td>
<td>-0.020</td>
<td>-18.915</td>
</tr>
<tr>
<td>nat.</td>
<td>-0.182</td>
<td>-0.017</td>
<td>-48.315</td>
</tr>
<tr>
<td>unnat.</td>
<td>-0.127</td>
<td>-0.017</td>
<td>-31.901</td>
</tr>
<tr>
<td>maj. prenom.</td>
<td>-0.109</td>
<td>-0.023</td>
<td>-7.278</td>
</tr>
<tr>
<td>maj. postnom.</td>
<td>-0.152</td>
<td>-0.023</td>
<td>-4.811</td>
</tr>
</tbody>
</table>

To examine if either of these conditioning contexts were used by participants we performed two different calculations of conditional entropy. The formula we used for conditional entropy looked as follows:

\[
H(V|C) = -\sum_{c_j \in C} p(c_j) \sum_{v_i \in V} p(v_i|c_j) \log_2 p(v_i|c_j)
\]

In one set of calculations the contexts \(C\) were the two head nouns that participants were exposed to, and in the other it was the two dependents (one genitive and one adjective meaning). However, simply calculating conditional entropy does not distinguish between participants who harmonise versus participants who employ lexical conditioning; becoming more consistent across the full system (harmony) also entails becoming more consistent within a specific context. To disentangle these strategies we combined the conditional entropy measures with the overall measures of entropy to calculate Mutual Information. Mutual Information (MI) for dependent type was calculate as:

\[
MI = \text{overall entropy} - \text{conditional entropy (dependent)}
\]

whereas Mutual Information for head noun was calculated as:

\[
MI = \text{overall entropy} - \text{conditional entropy (noun)}
\]

MI can take any value between 0 and 1, with 1 representing a system where the two variants (gesture orders) are perfectly conditioned on the context and 0 representing a system where knowing the context does not help you predict the variant used. As with the entropy calculation, the measures in Figure 4.5.7 and Figure 4.5.8 show mean change in MI (i.e. output MI - input MI) across condi-
Confidence intervals were generated in the same way as for the harmony calculations and, again, based on the same Monte Carlo simulation we extracted change in MI based on both measures of conditional entropy. We used z-scores to compare the experimental and simulated mean change in MI based on dependent type (Table 4.5.3) and head noun (Table 4.5.4). These indicated a significant change for nouns across all conditions, and MI change based on dependent type in all conditions except the the unnatural condition.

![Graph of MI change for different conditions](image)

**Figure 4.5.7:** Mean change in MI based on dependent type, showing reliable change in MI across all conditions except the unnatural condition in Experiment 1. Right-hand facet shows grand mean change in MI. Error bars show bootstrapped 95% CIs.

**Table 4.5.3:** Experimental means, simulated means and resulting z-scores for change in MI based on dependent type in Experiment 1. Z-scores show that experimental means are reliably different from the simulated means (indicating that MI increased significantly) in all conditions except the unnatural condition.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Exp. mean</th>
<th>Sim. mean</th>
<th>z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.099</td>
<td>0.028</td>
<td>10.583</td>
</tr>
<tr>
<td>natural</td>
<td>0.095</td>
<td>0.031</td>
<td>3.461</td>
</tr>
<tr>
<td>unnatural</td>
<td>0.025</td>
<td>0.031</td>
<td>-0.288</td>
</tr>
<tr>
<td>prenom</td>
<td>0.176</td>
<td>0.025</td>
<td>30.385</td>
</tr>
<tr>
<td>postnom</td>
<td>0.103</td>
<td>0.025</td>
<td>15.425</td>
</tr>
</tbody>
</table>
Figure 4.5.8: Mean change in MI based on head nouns, showing reliable change in MI across all conditions in Experiment 1. Right-hand facet shows grand mean change in MI. Error bars show bootstrapped 95% CIs.

4.5.3 Discussion

The results of this experiment replicate the findings reported in Holtz, Culbertson, and Kirby (2023). Experiment 1 found evidence of system-wide biases: participants tended to harmonise the language system they were trained on, and/or establish patterns of lexical conditioning. Crucially, by contrast, there was no evidence for category-specific biases: participants did not choose natural, bias-congruent orders any more often than they saw them in training. These results therefore support the claim that category-specific biases for adjectives and genitives do not influence learning behaviour in tasks which are low on the Scale of Innovation.

One interesting difference from Holtz, Culbertson, and Kirby (2023) is worth noting. Here, we saw less influence of participants’ native language. While participants in the majority prenominal condition (where the majority training order matches the majority order for both dependent types in English) showed a higher tendency to select the majority orders in the experiment in Holtz, Culbertson, and Kirby (2023), no such difference was found in this experiment. This suggests that effects of native language may not be particularly robust in this type of task, in line with previous work using silent gesture to limit the influence of native language (Goldin-Meadow et al., 2008).

To summarise, the results from Experiment 1 support the notion that tasks
Table 4.5.4: Experimental means, simulated means and resulting Z-scores for change in MI based on head noun in Experiment 1. Z-scores show that experimental means are reliably different from the simulated means (indicating that MI increased significantly) in all conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Exp. mean</th>
<th>Sim. mean</th>
<th>z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.055</td>
<td>0.021</td>
<td>15.785</td>
</tr>
<tr>
<td>nat.</td>
<td>0.069</td>
<td>0.018</td>
<td>14.344</td>
</tr>
<tr>
<td>unnat.</td>
<td>0.063</td>
<td>0.018</td>
<td>12.835</td>
</tr>
<tr>
<td>maj. prenom.</td>
<td>0.045</td>
<td>0.025</td>
<td>3.942</td>
</tr>
<tr>
<td>maj. postnom.</td>
<td>0.045</td>
<td>0.025</td>
<td>4.016</td>
</tr>
</tbody>
</table>

at the far left of the Scale of Innovation are most likely to activate system-wide biases for harmony and regularity, rather than category-specific biases that prefer postnominal adjectives and prenominal genitives.

### 4.6 Experiment 2

In Experiment 2, we examine whether a different type of task is more likely to activate the category-specific biases for postnominal adjectives and prenominal genitives. Specifically, the task in Experiment 2 requires participants to generalise from seen items to novel items of the same category at test (i.e., within category extension), while maintaining the same type of variable training data. This design utilises a combination of methodologies, meaning that the task in this experiment sits at the boundary of the regularising and generalising part of the Scale of Innovation. We know from previous work that category-specific ordering preferences for adjectives and genitives emerge when participants are not trained on any prior stimuli (Holtz, Kirby, & Culbertson, 2023). Our goal with this task is to reduce the evidence they have and therefore introduce the potential to innovate; the lack of direct evidence for these novel items might cause participants to fall back on their category-specific preferences when faced with uncertainty. However, novel items which are similar to (i.e., share a category with) previous items do not necessarily bring this about, it is an empirical question. Another possibility is that novel items which fit clearly into a system instead highlight the potential for systematic organisation of the language. This could lead to a boost in harmony rather than category-specific biases.
4.6.1 Method

The methods, materials and general procedure during training was identical to Experiment 1, but with some specific changes to instructions and the stimuli presented during the testing phase.

Materials

Identical to Experiment 1.

Procedure

Instead of being shown the same images as the ones they had seen during training, participants in Experiment 2 were given (only) new meanings during the testing phase. These meanings were comprised of the two nouns which participants had not been exposed to during training, combined with one new item from each dependent type. Prior to the testing phase participants were told “In this second part of the experiment, you will be shown a similar collection of images again as well as two gesture videos. Please click on the gesture videos that you think best conveys the meaning in the highlighted image.” Participants also answered the same demographics questions as in the previous experiment.

Participants

222 participants were recruited using the same methods and restrictions as in Experiment 1. The average estimated completion time for the experiment remained the same, as did the participant payment. Following preregistered exclusion criteria we excluded nine participants who reported proficiency in at least one sign language, as well as four participants who gave incoherent responses to the survey questions. A further two participants were excluded for consistently selecting the same button during the testing phase. Finally, five participants’ data was lost due to connection issues during study completion. The data in the final analysis comes from 52 participants in the natural condition, 50 in the unnatural condition, 46 in the majority prenominal condition, and 52 in the majority postnominal condition.

4.6.2 Results

If the introduction of novel items in this study increases the potential for innovation, then we predict that participants will show some evidence for category-specific biases, in contrast to Experiment 1. Specifically, we predict that partic-
Participants’ use of the majority order they were trained on will differ based whether that order is natural or not (i.e., whether the order aligns with the typological tendency, prenominal for genitives; postnominal for adjectives). This would lead to an overall increase in natural orders between input and output, and therefore a weaker tendency to harmonise. If, on the other hand, participants treat the appearance of novel items in the testing phase as a “complete the system” task, there is a different set of predictions. In this case, we may fail to see any effect of naturalness of the majority order. Rather, the overall tendency to harmonise may be stronger. We first analyse the results from Experiment 2 on their own, following the same process as in Experiment 1. We then compare them with the results from Experiment 1.

Learning

An overview of the output systems generated by participants in all conditions in Experiment 2 can be seen in Figure 4.6.1. We start by evaluating participants’ learning behaviour, that is, whether they chose gesture videos at test that matched the majority order they were trained on. The model used to evaluate learning behaviour was identical to the one used in Experiment 1, with a binary outcome variable for majority order and fixed effects for condition and dependent type. The same random effect structure also applied. The model had a significant positive intercept ($\beta = 0.73, SE = 0.14, z = 5.05, p < 0.001$), suggesting that participants across all conditions, on average, chose the majority order they were trained on at a rate above chance. This trend is also visually present in Figure 4.6.2, which plots proportion of majority orders chosen by participants during the testing phase, split by dependent type. Model comparison using a likelihood ratio test found that the model including an interaction between condition and dependent type improved model fit ($\chi^2 = 8.38, p = 0.039$). In this model there was a negative effect in the majority postnominal condition such that participants selected fewer majority orders for genitives in this condition, compared to the grand mean ($\beta = -1.08, SE = 0.52, z = -2.06, p = 0.04$). No other effects were significant. This shows that learning was mostly comparable across all conditions, with the exception of genitives in the majority postnominal condition.

Naturalness

Next, we analysed participants’ tendency to choose natural orders. Figure 4.6.3 shows the individual and overall proportions of natural orders that participants selected at test, split by dependent type. We ran a model with a binary outcome
variable, *natural*, with the same structure as above. The model had a significant positive intercept ($\beta = 0.42$, $SE = 0.15$, $z = 2.74$, $p < 0.01$), suggesting that participants across all conditions, on average, chose the natural order at a rate above chance. Model comparison found that a model with an interaction between *condition* and *dependent type* had the best fit ($\chi^2 = 12.05$, $p < 0.01$). There was a significant positive effect for *condition*, such that participants in the natural condition selected more natural orders at test, compared to the grand mean ($\beta = 0.67$, $SE = 0.26$, $z = 2.56$, $p = 0.01$). Similarly, there was a significant positive effect for *dependent type*, such that participants selected more natural orders for genitives, compared to the grand mean ($\beta = 0.88$, $SE = 0.29$, $z = 3.03$, $p < 0.01$).

In addition to these positive effects there was also a negative effect of *condition* for the unnatural condition, such that participants in this condition selected less natural orders at test ($\beta = -0.81$, $SE = 0.26$, $z = -3.09$, $p < 0.01$). Finally, as in Experiment 1, the interaction between *condition* and *dependent type* was significant for both majority conditions. The direction of the interaction effect depended on if the majority order of the condition matched the natural order for a given dependent. For the majority postnominal condition the effect was negative, as participants chose the natural order for genitives less often, compared to the grand mean ($\beta = -1.40$, $SE = 0.50$, $z = -2.82$, $p < 0.01$). Whereas the effect was positive in the majorityprenominal condition, as participants chose the natural
order more often, compared to the grand mean ($\beta = 1.81, SE = 0.52, z = 2.90, p < 0.01$). Again, these results are mainly a reflection of the overall learning behaviour. As participants learned the orders they were trained on they selected more natural orders when the majority order was natural (either for a specific dependent type or across the full condition), and selected less natural orders when the majority order was unnatural (either for a specific dependent type or across the full condition), compared to the grand mean. The main difference between these results and those from Experiment 1 is in the significant positive intercept, showing that participants in Experiment 2 selected more natural orders than is expected by chance.

**Harmony**

Participants’ tendency to harmonise was evaluated in the same way as for Experiment 1. After calculating the mean change in entropy for each condition, as well as the overall change (see Figure 4.6.4), we performed the same bootstrapping analysis as before. Based on this we found that entropy change was reliable in the natural and unnatural condition, but not in the two majority order conditions. To verify the reliability of these changes we calculated z-scores between experimental and simulated means (see Table 4.6.1) based on the same kind of Monte
Figure 4.6.3: Proportion of test trials where participants selected the natural order (i.e., prenominal for genitives, postnominal for adjectives) for each condition and dependent type in Experiment 2. Right-hand facet shows the grand mean alongside the mean for Experiment 1 for comparison. There was an overall preference for natural orders. Error bars show bootstrapped 95% CIs.

Carlo simulation as in Experiment 1.\textsuperscript{8} The result of this comparison showed that change in entropy was reliable in all conditions except the majority postnominal condition, despite the fact that the CIs for the majority prenominal order crosses zero in Figure 4.6.4.

Table 4.6.1: Experimental means, simulated means and resulting z-scores for change in overall entropy in Experiment 2. Z-scores show that experimental means are reliably different from the simulated means (indicating that entropy dropped significantly) in all conditions except the majority postnominal condition.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Exp. mean</th>
<th>Sim. mean</th>
<th>z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>-0.142</td>
<td>-0.020</td>
<td>-18.601</td>
</tr>
<tr>
<td>nat.</td>
<td>-0.264</td>
<td>-0.017</td>
<td>-72.339</td>
</tr>
<tr>
<td>unnat.</td>
<td>-0.202</td>
<td>-0.017</td>
<td>-53.788</td>
</tr>
<tr>
<td>maj. prenom.</td>
<td>-0.075</td>
<td>-0.023</td>
<td>-2.932</td>
</tr>
<tr>
<td>maj. postnom.</td>
<td>-0.014</td>
<td>-0.023</td>
<td>0.465</td>
</tr>
</tbody>
</table>

\textsuperscript{8}Note that the simulated means for these experiments are identical as the change in testing items did not change the probabilities in the model.
Figure 4.6.4: Mean change in entropy between input and output, showing participants tend to harmonise in the natural and unnatural condition, but not in the majority prenominal or majority postnominal condition in Experiment 2. Right-hand facet shows grand mean change in entropy alongside the mean change for experiment 1 for comparison. Error bars show bootstrapped 95% CIs.

**Conditioned variation**

Although we did not have any clear prediction about lexically conditioned variation based on our manipulation of the task, we still look for evidence for this kind of strategy. Despite the fact that both the nouns and dependents are novel at test in Experiment 2, it is still possible for participants to decide to use one gesture order for each dependent type, or one order for each head noun. However, here these strategies cannot be attached to a previously seen lexical item. We evaluated the presence of conditioning context using the same calculations of conditional entropy and change in Mutual Information between input and output (see Figure 4.6.5 for MI change based on dependent type and Figure 4.6.6 for MI change based on head noun). Confidence intervals were generated in the same way as in Experiment 1 and, again, we verified these calculations using the same kind of Monte Carlo simulation as before. We extracted change in MI based on both measures of conditional entropy from the simulation and compared these to the experimental means using z-scores. Results of these comparisons can be seen in Table 4.6.2 for dependent type and Table 4.6.3 for head noun. We find that change in MI was significant for both measures across all conditions, even though some of the CIs in Figure 4.6.5 cross zero. This analysis shows that participants in Experiment 2 employed both dependent type and individual nouns as condi-
tioning contexts for the two gesture orders as a way to reduce the unpredictable variation that was present in their training data.

![Graph showing change in MI for dependents across different conditions](image)

Figure 4.6.5: Mean change in MI based on dependent type in Experiment 2, showing reliable change in MI for the two majority order conditions, but not the natural or unnatural condition. Right-hand facet shows grand mean change in MI alongside the mean change for Experiment 1 for comparison. Error bars show bootstrapped 95% CIs.

Table 4.6.2: Experimental means, simulated means and resulting Z-scores for change in MI based on dependent type in Experiment 2. Z-scores show that experimental means are reliably different from the simulated means (indicating that MI increased significantly) in all conditions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Exp. mean</th>
<th>Sim. mean</th>
<th>Z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.178</td>
<td>0.028</td>
<td>22.165</td>
</tr>
<tr>
<td>nat.</td>
<td>0.066</td>
<td>0.031</td>
<td>1.887</td>
</tr>
<tr>
<td>unnat.</td>
<td>0.070</td>
<td>0.031</td>
<td>2.093</td>
</tr>
<tr>
<td>maj. prenom.</td>
<td>0.270</td>
<td>0.025</td>
<td>49.353</td>
</tr>
<tr>
<td>maj. postnom.</td>
<td>0.312</td>
<td>0.025</td>
<td>56.514</td>
</tr>
</tbody>
</table>

4.6.3 Discussion

Experiment 2 examined whether having novel items which required within-category extension at test would lead to the emergence of category-specific ordering biases for adjective and genitive meanings. This was grounded in the idea that, without
direct evidence, participants may instead choose to order these novel meanings based on their category-specific order preferences. We know from previous work that these ordering preferences emerge when participants are not trained on any prior stimuli (Holtz, Kirby, & Culbertson, 2023). Our goal was to reduce the evidence participants could use to make their judgements, and therefore encourage them to fall back on their category-specific preferences. However, we also noted that the presence of novel items may instead highlight the existence of a system. This might lead participants to view the testing phase as a “complete the system” type of task. Thus we also entertained the idea that this could boost the influence of some system-wide biases on participants’ behaviour.

The results of Experiment 2 show that, as in Experiment 1, participants learned from the training data they were given. Similarly, there was an overall tendency to harmonise the language at test, but this tendency was not significant for the majority postnominal condition. Participants in both majority prenominal and postnominal conditions seemed to search for conditioning contexts for the variation they had observed, reflected in the especially high increase in MI based on dependents for these two conditions, although the increase was also significant in the other two conditions. The MI calculations based on noun suggests that these remained available as a conditioning context to participants in Experiment 2, despite the fact that the nouns themselves changed between training and
testing. We return to this pattern in the general discussion.

Crucially, in Experiment 2 we also observed an effect of naturalness. While there was no difference in participants’ likelihood of choosing the majority training order based on naturalness, there was a significant increase in the number of natural orders that participants chose. This result is different from Experiment 1, where no such effect was found. This overall increase in natural orders suggests that participants in Experiment 2 were (at least to some extent) influenced by category-specific word order preferences for adjectives and genitives.

### 4.7 Experiment comparison

The individual analyses of the data from Experiment 1 and 2 show that, in both experiments, we see evidence that participants learn the majority orders they were trained on. We also saw strategies for reducing the unconditioned variation, including harmony and the introduction of conditioning contexts for the two word order variants based on dependent type and/or nouns. A key difference between the two experiments is in naturalness: there was no evidence for the role of category-specific biases in Experiment 1, but some evidence for this in Experiment 2. Below, we test if this difference reflects a difference in learning of natural orders between the two experiments.

In order to allow for easy comparison between the data from both experiments we created three new variables. The first variable, *experiment*, has one level for Experiment 1 (exp1) and one level for Experiment 2 (exp2). Next, we categorise the data based on two things: the naturalness status of the majority order that participants saw during training for a given dependent type; and the harmonic status of the training language. The *nat_unnat* variable thus had two levels,
unnat when the majority order was unnatural and nat when the majority order
was natural. Similarly, harm_non was harm when the training language was
mainly harmonic and non when it was mainly non-harmonic. The way this coding
scheme mapped on to the data can be seen in Table 4.7.1.

Table 4.7.1: Mapping of coding for variables nat_unnat and harm_non across
both Experiments 1 and 2 for experiment comparison.

<table>
<thead>
<tr>
<th>Condition</th>
<th>nat</th>
<th>unnat</th>
<th>maj. prenom</th>
<th>maj. postnom</th>
</tr>
</thead>
<tbody>
<tr>
<td>dependent</td>
<td>adj</td>
<td>gen</td>
<td>adj</td>
<td>gen</td>
</tr>
<tr>
<td>nat_unnat</td>
<td>nat</td>
<td>nat</td>
<td>unnat</td>
<td>unnat</td>
</tr>
<tr>
<td>harm_non</td>
<td>non</td>
<td>non</td>
<td>non</td>
<td>non</td>
</tr>
</tbody>
</table>

This allowed us to evaluate the triple interaction between experiment, nat_unnat,
and harm_non with majority order as the binary outcome variable. All fixed ef-
cfects were deviation coded (0.5, -0.5). Model comparison using a likelihood ratio
test showed that the model with the triple interaction was the best fit \( \chi^2 = 19.87, p < 0.001 \). The model had a significant positive intercept \( \beta = 0.74, SE = 0.08, z = 9.76, p < 0.001 \), reflecting the fact that, overall participants selected
more majority orders than expected by chance, i.e., they learned from their input.
Additionally, significant positive coefficients were also observed individually for
nat_unnat and experiment (nat_unnat: \( \beta = 0.33, SE = 0.11, z = 2.98, p < 0.01 \);
experiment: \( \beta = 0.49, SE = 0.15, z = 3.24, p < 0.01 \)). No other effects were sig-
nificant. These results show that, across both experiments together, participants
selected more majority orders for a given dependent type when the majority order
they had been trained on was natural, and that participants generally selected
more majority orders in Experiment 1, compared to the grand mean.

Collectively these results tell us that participants were less likely to choose the
orders they were trained on at test in Experiment 2. However across both exper-
iments, when the orders they were trained was natural, they were more likely to
choose them at test. The lack of an interaction between nat_unnat and experiment suggests that, contrary to what the analysis of the experiments individually
suggests, the generalisation task used in Experiment 2 did not necessarily lead to
a greater learning boost for natural orders. We return to this point below in the
general discussion.
4.8 General discussion

This main aim of this paper has been to examine whether category-specific biases can emerge in a broader range of contexts than previously seen. These biases have been mainly found in contexts involving high levels of improvisation—for example tasks where participants must improvise or judge gesture sequences without any prior evidence. And yet, these kinds of biases appear to have shaped language typology—patterns which correspond to them are over-represented the world’s languages. This is surprising, as, intuitively this type of improvisation would seem to be restricted to contexts of language emergence. This contrasts with system-wide biases, which have been found to emerge during learning, and thus have the potential to iteratively influence language structure at each new generation. Here, we explored the possibility that category-specific biases might emerge so long as there is some amount of innovation required in the task. We set out how innovation fits into different natural language and experimental contexts using the Scale of Innovation (recall Figure 4.4.1). On this scale, tasks requiring only or mainly memorisation are lowest, and tasks requiring only or mainly improvisation are highest. However, in the middle of the scale are contexts involving extrapolation or generalisation. There require extension of a system beyond direct evidence.

We hypothesised that the amount of innovation involved in a linguistic task influences the probability of observing the effects of category-specific cognitive biases. We tested this in the domain of nominal word order, where previous research has shown that there is a postnominal preference for adjectives and a prenominal preference for genitives in contexts of high innovation (Culbertson, Schouwstra, & Kirby, 2020; Holtz, Kirby, & Culbertson, 2023; Jaffan et al., 2020). If these were the only contexts in which category-specific biases influence people’s word order preferences then we would not expect to see them reflected in the typological data; we would instead expect them to be swamped by system-wide biases which are active during learning, like harmony (Culbertson, 2012; Holtz, Culbertson, & Kirby, 2023). But both appear to be reflected in the typology of these two dependent types.

To test our hypothesis, we compared a low-innovation context — an artificial language learning task with a regularisation design and only seen items at test (Experiment 1) — to a higher-innovation task which required generalisation to novel items at test (Experiment 2). We found that, while participants in Experiment 1 only showed evidence of being influenced by system-wide biases for harmony and regularity, participants in Experiment 2 were also showed some
evidence of being influenced by category-specific biases for adjectives and gen-

tives; they selected more natural orders at test than was expected by chance.

This suggests that the introduction of new items in the testing phase of the gen-

eralisation experiment was enough to (at least partially) activate participants’

category-specific ordering preferences in the absence of direct evidence for how
	hese items fitted into a wider language system. These results support the idea

that, like natural language learning and use, experimental tasks exist along a

Scale of Innovation, since manipulating how much of the system participants has

access to affected their behaviour.

While we observed some effects of category-specific biases across experiments,

their effects did not emerge in quite the way we expected. We predicted that, if

participants viewed the task in Experiment 2 as requiring more innovation, then

their learning behaviour—i.e., the degree to which they chose the trained order

at test—should be influenced by the naturalness of the items they were trained

on in that experiment. This is not exactly what we observed. Our comparison

between the experiments revealed no significant interaction between nat_unnat

and experiment, i.e. participants in Experiment 2 did not choose natural majority

orders more than unnatural orders for either dependent type compared to Experiment 1. Our prediction here was grounded in the results from other

experiments that have found some effect of category-specific ordering biases in

learning behaviour for other word order patterns (Do et al., 2022; Motamedi,

Wolters, Schouwstra, & Kirby, 2022). In these studies there was some evidence

that natural/preferred orders were also learned better. We also find some support

for this same pattern in the overall effect of naturalness on learning across the

two experiments (i.e. in the main effect of nat_unnat). However, this effect did

not differ across experiments.

Participants in both experiments exhibited clear evidence of some system-wide

biases. In Experiment 1 the bias for harmony was found consistently across all

four conditions, and both nouns and dependents tended to be used as conditioning

strategies for reducing unpredictable variation. The harmonic tendency in Ex-

periment 2 was more varied, with all conditions except the majority postnominal

condition showing a significant increase in harmony when compared against sim-

ulated results. Yet participants in Experiment 2 also used both dependent type

and head nouns as conditioning contexts. The harmony results in Experiment

1 are in line with previous research which has identified a harmony preference

in both ease of learning paradigms and regularisation paradigms (Christiansen,

2000; Culbertson & Newport, 2015; Culbertson et al., 2012). The inconsistency

of those same results across conditions in Experiment 2 speaks to the potential
competition between the system-wide bias for harmony and the category-specific biases keeping the dependents split across the head noun.

It is interesting that both dependent type and head noun seemed to be used as conditioning contexts across both experiments, given the fact that the strategies are arguably less accessible in Experiment 2. In Experiment 1, the conditioning context could be tied to specific words for both dependents and nouns, since the same items were used in training and testing. In contrast, Experiment 2 had different nouns and dependents in the testing phase. While participants could still have arbitrarily assigned one order to a given noun or dependent, this strategy differs from assigning it to a specific previously seen word, as is possible in Experiment 1. This difference between the training and testing items could have resulted in participants being unable to use dependents and nouns as conditioning contexts in Experiment 2, but generally both conditioning strategies were also used in Experiment 2. This shows that participants in Experiment 2 were able to apply conditioning contexts to reduce unpredictable variation, even when the items in those contexts were novel. Consequently, in addition to the system-wide preference for harmony being active in both experiments, the results from both experiments also follow previous research showing that people also tend to reduce unpredictable variation by imposing conditioning contexts for the variation (Samara et al., 2017; Smith & Wonnacott, 2010).

In contrast to the results of previous studies, the experiments in this study showed less influence from native language on participants’ learning and regularisation behaviour. In previous learning studies looking at similar meanings, participants in the majority prenominal condition showed significantly better learning than participants in the other conditions (Holtz, Culbertson, & Kirby, 2023). This was attributed to a learning boost from English in this specific condition, since the majority order in the prenominal condition matches the default order for adjectives and genitives in English. The lack of such an effect in the results of both experiments in this study suggests that gestural stimuli, while not always eliminating native language influence, can act to reduce native-speaker effects in experimental work (Goldin-Meadow et al., 2008).

Although a direct comparison between experiments did not support a significant difference in learning based on naturalness, we nevertheless saw an individual preference for natural orders in Experiment 2 which was not present in Experiment 1. Our expectation is that if the semantic distance between the training items and the testing items is increased, this would increase the need for innovation in the testing phase and result in a bigger influence of category-specific biases on participants’ behaviour. We would thus expect to see a stronger preference
for postnominal adjectives and prenominal genitives in such a task, compared to what we have seen here. One potential way to examine this would be to have them generalise from pattern adjectives to colour adjectives, and from human to non-human possessors (note that this type of possessive relationship is very different from the one we associate with animate possessors as it usually involves part-whole relations). Taking this a step further, training could also be conducted on one category of dependents and tested on a different category of dependents, taking the task from a generalisation to an extrapolation context. For example, training participants on the order between adjectives and nouns and then testing them on the order of genitives and nouns, similar to what is done in Wang et al. (2023) where participants extrapolate the order between adjectives and nouns to the order between verbs and objects (and vice versa). This would allow for further examination of exactly where along the Scale of Innovation that category-specific and system-wide biases have an effect, since the extrapolation context has not been targeted for this set of opposing cognitive biases. Given that extrapolation contexts arguably involve more innovation than generalisation tasks, we expect that the tentative influence of category-specific biases in the experiments we presented in this study would be stronger in an extrapolation context.

4.9 Conclusion

This study introduced the notion that linguistic contexts can be ordered along a Scale of Innovation, based on the amount of innovation that is required to perform the linguistic tasks in each type of context. Based on previous research on the effect of cognitive biases on linguistic structure, we discussed two types of cognitive biases, system-wide and category-specific. We proposed that the effects of these biases are strongest in tasks at different ends of the Scale of Innovation. System-wide biases seem to be active during low-innovation contexts that mainly rely on system learning, and their cumulative effects on language are large, whereas category-specific biases tend to be observed in high-innovation contexts involving language creation and improvisation, which are rare in the present-day language environment. However, this study found evidence suggesting that category-specific biases may also have some influence on individuals’ behaviour in contexts that require a mixture of learning and innovation. This was mainly found in the fact that participants in Experiment 2 chose more natural orders than what would be expected by chance. Additionally, the learning effect, whereby participants learned orders which followed category-specific pref-
erences more readily across both experiments, suggests that these biases may have some limited effect on behaviour even in contexts as far along the scale as *regularisation*. This evidence suggests that category-specific biases may have a wider range of opportunity to influence language structure than previously thought. This in turn helps to explain why we can still see them reflected in typological data. While the small increase in the amount of innovation required in the generalisation task did not affect participants’ behaviour in the comparison between experiments, future research that manipulates the innovation required in the experimental task will help to develop our understanding of how different cognitive biases shape language across contexts.
References


228
Dryer, M. S. (2013a). Order of adjective and noun. In M. S. Dryer & M. Haspelmath (Eds.), The world atlas of language structures online. Max Planck Institute for Evolutionary Anthropology.

Dryer, M. S. (2013b). Order of genitive and noun. In M. S. Dryer & M. Haspelmath (Eds.), The world atlas of language structures online. Max Planck Institute for Evolutionary Anthropology.


231


Wang, F., Kirby, S., & Culbertson, J. (2023). The learning bias for cross-category harmony is sensitive to semantic similarity: Evidence from artificial lan-

Appendix

4.A Outline

This appendix includes further information about additional data collected in the experiments from Chapter 4. Sections 4.B and 4.C give details about how informed consent was obtained from participants, and the instructions they saw during the experiments. Following this 4.D shows full tables where CIs for change in overall entropy are compared between conditions for both experiments. 4.E gives details about the spoken languages that participants reported knowing when they answered the questionnaire. Finally, section 4.F describes the set-up and results of the Monte Carlo simulations that were used to evaluate the reliability of the changes in entropy and Mutual Information that were observed in the experiments.

4.B Consent info

The experiments included a web-page where participants could indicate their consent to take part in the study by ticking a box and pressing a button to progress the experiment. This page also included the information sheet, with specific information about data storage, right to withdraw etc. Below you will find all the text included on this web-page for Experiment 1 (first page) and Experiment 2 (second page).
Study title: Gesture Study 1

Principal Investigators: Jennifer Culbertson, Simon Kirby
Researcher collecting data: Annie Holtz

What is this document? This document explains what kind of study we’re doing, what your rights are, and what will be done with your data. You should print this page for your records.

Nature of the study. You are invited to participate in a study that involves observing videos showing a person using gesture in a made-up sign language to convey information. You will see stimulus pictures accompanied by one or two videos showing gestures, and be asked to make a choice between them. Once you finish, we may also have some questions about your experience (e.g., age, language background). Your responses will be recorded. Your session should last for up to 10 minutes. You will be given full instructions shortly.

Compensation. You will be paid £1.74 for your participation in this study.

Risks and benefits. There are no known risks to participation in this study. Other than the payment mentioned, there are no tangible benefits to you, however you will be contributing to our knowledge about language.

Confidentiality and use of data. All the information we collect during the course of the research will be processed in accordance with Data Protection Law. In order to safeguard your privacy, we will never share personal information (like your name) with anyone outside the research team. Your data will be referred to by a unique participant number rather than by name. Please note that we will temporarily collect your Prolific ID to prevent repeat participation; however, we will never share this information with anyone outside the research team and it will be deleted two weeks after data collection. We will store any personal data (i.e., Prolific ID) using the University of Edinburgh’s secure encrypted storage service. The anonymised data collected during this study will be used for research purposes.

What are my data protection rights? The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner’s Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Voluntary participation and right to withdraw. Your participation is voluntary, and you may withdraw from the study at any time and for any reason. If you withdraw from the study during data gathering, we will delete your data and there is no penalty or loss of benefits to which you are otherwise entitled. If you wish to withdraw after you have completed the study we will only be able to delete your data, or to remove it from inclusion in published results for two weeks after the date of your participation. Once again, withdrawing at this stage will not cause any penalty or loss of benefits to which you are otherwise entitled. If you have any questions about what you’ve just read, please feel free to ask, or contact us later. You can contact us by email at ... This project has been approved by PPLS Ethics committee. If you have questions or comments regarding your rights as a participant, they can be contacted at 0131 650 4020 or ppls.ethics@ed.ac.uk.

By taking part in the experiment, you consent to the following:
1. I agree to participate in this study.
2. I confirm that I have read and understood how my data will be stored and used.
3. I understand that I have the right to terminate this session at any point. If I choose to withdraw after completing the study, and request for my data to be removed within two weeks of my participation, my data will be deleted.

To continue, tick the checkbox below and click "Start Experiment".

☐ I agree to take part in this study.

Start Experiment
Study title: Gesture Study 2

Principal Investigators: Jennifer Culbertson, Simon Kirby
Researcher collecting data: Annie Holtz

What is this document? This document explains what kind of study we’re doing, what your rights are, and what will be done with your data. You should print this page for your records.

Nature of the study. You are invited to participate in a study that involves observing videos showing a person using gesture in a made-up sign language to convey information. You will see stimulus pictures accompanied by one or two videos showing gestures, and be asked to make a choice between them. Once you finish, we may also have some questions about your experience (e.g., age, language background). Your responses will be recorded. Your session should last for up to 10 minutes. You will be given full instructions shortly.

Compensation. You will be paid £1.74 for your participation in this study.

Risks and benefits. There are no known risks to participation in this study. Other than the payment mentioned, there are no tangible benefits to you, however you will be contributing to our knowledge about language.

Confidentiality and use of data. All the information we collect during the course of the research will be processed in accordance with Data Protection Law. In order to safeguard your privacy, we will never share personal information (like your name) with anyone outside the research team. Your data will be referred to by a unique participant number rather than by name. Please note that we will temporarily collect your Prolific ID to prevent repeat participation, however we will never share this information with anyone outside the research team and it will be deleted two weeks after data collection. We will store any personal data (i.e., Prolific ID) using the University of Edinburgh’s secure encrypted storage service. The anonymised data collected during this study will be used for research purposes.

What are my data protection rights? The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner’s Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Voluntary participation and right to withdraw. Your participation is voluntary, and you may withdraw from the study at any time and for any reason. If you withdraw from the study during data gathering, we will delete your data and there is no penalty or loss of benefits to which you are otherwise entitled. If you wish to withdraw after you have completed the study we will only be able to delete your data, or to remove it from inclusion in published results for two weeks after the date of your participation. Once again, withdrawing at this stage will not cause any penalty or loss of benefits to which you are otherwise entitled.

If you have any questions about what you’ve just read, please feel free to ask, or contact us later. You can contact us by email at [email protected]. This project has been approved by PPLS Ethics committee. If you have questions or comments regarding your rights as a participant, they can be contacted at 0131 650 4020 or pplsethics@ed.ac.uk.

By taking part in the experiment, you consent to the following:
1. I agree to participate in this study.
2. I confirm that I have read and understood how my data will be stored and used.
3. I understand that I have the right to terminate this session at any point. If I choose to withdraw after completing the study, and request for my data to be removed within two weeks of my participation, my data will be deleted.

To continue, tick the checkbox below and click "Start Experiment".

☐ I agree to take part in this study.

Start Experiment
4.C Instructions

4.C.1 Experiment 1: Instructions

The instructions were identical across all four conditions in the experiment. Below are all the instructions from each segment of the experiment.

Contrast trial instructions: Instructions provided prior to the contrast/familiarisation trials across all conditions.

In the first part of the experiment, you will observe gestures being produced which correspond to the meaning of different images. The gestures show how you express ownership of an item or how you describe an item in a made-up sign language. On the next two screens you will see examples of the kinds of images you will see in the rest of the experiment. The images vary by the item they show, the owner and the pattern. Study the images and then continue to the next screen.

Training instructions: Instructions provided before the training phase across all conditions.

In the first part of the experiment, one of the images in the grid will be highlighted and underneath the images you will see a video of a person using gestures to convey the meaning of the highlighted image. These kinds of images, along with their accompanying gestures, will be shown to you several times. Just sit back and watch carefully.

Testing instructions: Instructions provided before the testing phase across all conditions.

In this second part of the experiment, you will be shown the same collection of images again as well as two gesture videos. Please click on the corresponding gesture video like you saw in the first part. This will repeat several times for each image.

Questionnaire prompts: Instructions and prompts presented during the questionnaire across all conditions.

You have completed the testing phase. You will now be asked two quick questions about your language background and then you’re done!

Please answer the following questions:
1. Do you have any knowledge of a sign language? (If yes, please state which).

2. What spoken languages do you know and to what proficiency? Please indicate proficiency on a scale of 1-10 (e.g. English - 10, Tamil - 8, French - 4).

4.C.2 Experiment 2: Instructions

The instructions were identical across all four conditions in the experiment.
Contrast trial instructions: Identical to Experiment 1.
Training instructions: Identical to Experiment 1.
Testing instructions: Instructions provided before the testing phase across all conditions.

In this second part of the experiment, you will be shown a similar collection of images again as well as two gesture videos. Please click on the gesture videos that you think best conveys the meaning in the highlighted image. This will repeat several times for each image.

Questionnaire prompts: Identical to Experiment 1.

4.D Entropy CIs comparison

Table 4.D.1: Comparison of overall mean entropy change between conditions for Experiment 1 in Chapter 4. Includes 95% bootstrapped CIs around each mean. All confidence intervals cross zero, suggesting that no difference between conditions is significant.

<table>
<thead>
<tr>
<th>Condition comparison</th>
<th>$\overline{\mu}_a - \overline{\mu}_b$</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>nat. – unnat.</td>
<td>-0.05</td>
<td>-0.16</td>
<td>0.03</td>
</tr>
<tr>
<td>nat. – maj.prenom.</td>
<td>-0.07</td>
<td>-0.21</td>
<td>0.06</td>
</tr>
<tr>
<td>nat. – maj.postnom</td>
<td>-0.03</td>
<td>-0.16</td>
<td>0.10</td>
</tr>
<tr>
<td>unnat. – maj.prenom.</td>
<td>-0.02</td>
<td>-0.13</td>
<td>0.11</td>
</tr>
<tr>
<td>unnat. – maj.postnom.</td>
<td>0.03</td>
<td>-0.08</td>
<td>0.15</td>
</tr>
<tr>
<td>maj.prenom. – maj.postnom</td>
<td>0.04</td>
<td>-0.11</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Table 4.D.2: Comparison of overall mean entropy change between conditions for Experiment 2 in Chapter 4. Includes 95% bootstrapped CIs around each mean. Some CIs do not cross zero indicating reliable differences between the natural condition and both majority conditions, and between the unnatural and majority postnominal condition.

<table>
<thead>
<tr>
<th>Condition comparison</th>
<th>$\bar{x}_a - \bar{x}_b$</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>nat. – unnat.</td>
<td>-0.06</td>
<td>-0.19</td>
<td>0.064</td>
</tr>
<tr>
<td>nat. – maj.prenom.</td>
<td>-0.19</td>
<td>-0.33</td>
<td>-0.05</td>
</tr>
<tr>
<td>nat. – maj.postnom</td>
<td>-0.25</td>
<td>-0.37</td>
<td>-0.12</td>
</tr>
<tr>
<td>unnat. – maj.prenom.</td>
<td>-0.13</td>
<td>-0.26</td>
<td>0.01</td>
</tr>
<tr>
<td>unnat. – maj.postnom.</td>
<td>-0.19</td>
<td>-0.31</td>
<td>-0.07</td>
</tr>
<tr>
<td>maj.prenom. – maj.postnom.</td>
<td>-0.06</td>
<td>-0.20</td>
<td>0.08</td>
</tr>
</tbody>
</table>

4.E Language data

Data in Tables 4.E.1 shows the spoken languages that participants in Experiment 1 in Chapter 4 reported having any knowledge of. The same data for Experiment 2 is found in Table 4.E.2. As participants in both experiments knew several languages the counts here exceeds the total number of participants. Languages are sorted from most frequent to least frequent, with alphabetical ordering for languages with the same number of speakers.

Table 4.E.1: Languages (other than English) spoken by participants whose data was included in Experiment 1 in Chapter 4.

<table>
<thead>
<tr>
<th>Languages</th>
<th>N participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>32</td>
</tr>
<tr>
<td>Spanish</td>
<td>22</td>
</tr>
<tr>
<td>German</td>
<td>21</td>
</tr>
<tr>
<td>Italian</td>
<td>7</td>
</tr>
<tr>
<td>Afrikaans</td>
<td>5</td>
</tr>
<tr>
<td>Zulu</td>
<td>5</td>
</tr>
<tr>
<td>Welsh</td>
<td>4</td>
</tr>
<tr>
<td>Mandarin</td>
<td>3</td>
</tr>
<tr>
<td>Thai</td>
<td>3</td>
</tr>
<tr>
<td>Xhosa</td>
<td>3</td>
</tr>
<tr>
<td>Cantonese</td>
<td>2</td>
</tr>
<tr>
<td>Dutch</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 4.E.2: Languages (other than English) spoken by participants whose data was included in Experiment 2 in Chapter 4.

<table>
<thead>
<tr>
<th>Languages</th>
<th>N participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>40</td>
</tr>
<tr>
<td>Spanish</td>
<td>18</td>
</tr>
<tr>
<td>German</td>
<td>14</td>
</tr>
<tr>
<td>Japanese</td>
<td>7</td>
</tr>
<tr>
<td>Italian</td>
<td>6</td>
</tr>
<tr>
<td>Mandarin</td>
<td>6</td>
</tr>
<tr>
<td>Cantonese</td>
<td>4</td>
</tr>
<tr>
<td>Irish</td>
<td>3</td>
</tr>
</tbody>
</table>
4.F Monte Carlo simulation

As in the previous two chapters I performed a Monte Carlo simulation to validate the reliability of the changes in overall entropy and and mutual information that were calculated based on the experimental data from both experiments in Chapter 4. This simulation was used to compare the observed experimental values for change in overall entropy and change in Mutual Information (MI) based on both head noun and dependent type, to a theoretical baseline of sampling based on input probability.

The simulation was identical to the one used in Chapter 3. The simulation ran for 10,000 runs through the experiment, with 50 agents (i.e. simulated partici-
pants) assigned to each condition. Both experiments in this chapter had 32 testing trials so each agent in the simulation sampled between the two variants (prenominal or postnominal order) 32 times, generating 64,000,000 (32 * 200 * 10,000) data points. The probabilities of sampling were the same as before, 0.75 for the majority order and 0.25 for the minority order.

The same procedure was used for both the experimental and simulated data to calculate mean change in overall entropy and mean change in MI based on head nouns and dependent type for each condition and across conditions for each run of the simulation. A comparison between the simulated means for change in overall entropy and change in MI for both head nouns and dependents can be seen in Figures 4.F.1-4.F.6 for Experiment 1 and Figures 4.F.7-4.F.12 for Experiment 2. As the z-score calculations for the experiment confirmed, all experimental values are significantly different from the simulated means except MI change based on dependent type for the the unnatural condition.

4.F.1 Experiment 1: Monte Carlo

![Simulated and experimental means of change in overall entropy](image)

Figure 4.F.1: Distribution based on simulated means of change in overall entropy across conditions, including visual comparison between simulated mean of means and observed experimental mean for Experiment 1 in Chapter 4. Experimental mean is significantly different from simulated mean of means.

243
Figure 4.F.2: Distribution based on simulated means of change in overall entropy for each condition, including visual comparison between simulated means of means and observed experimental means for Experiment 1 in Chapter 4. Experimental means are significantly different from simulated means of means for all conditions.

Figure 4.F.3: Distribution of simulated means of change in MI based on head noun across conditions, including visual comparison between simulated mean of means and observed experimental mean for Experiment 1 in Chapter 4. Experimental mean is significantly different from simulated mean of means.
Figure 4.F.4: Distribution of simulated means of change in MI based on head noun for each condition, including visual comparison between simulated means of means and observed experimental means for Experiment 1 in Chapter 4. Experimental means are significantly different from simulated means of means for all conditions.

Figure 4.F.5: Distribution of simulated means of change in MI based on dependent type across conditions, including visual comparison between simulated mean of means and observed experimental mean for Experiment 1 in Chapter 4. Experimental mean is significantly different from simulated mean of means.
Figure 4.F.6: Distribution of simulated means of change in MI based on dependent type for each condition, including visual comparison between simulated means of means and observed experimental means for Experiment 1 in Chapter 4. Experimental means are significantly different from simulated means of means for all conditions except the unnatural condition.
4.F.2 Experiment 2: Monte Carlo

The simulation for experiment 2 was identical to that of Experiment 1. Due to this, all the simulated means are the same as in the previous section. Differences are found in the comparison between the simulated and experimental means.

Figure 4.F.7: Distribution based on simulated means of change in overall entropy across conditions, including visual comparison between simulated mean of means and observed experimental mean for Experiment 2 in Chapter 4. Experimental mean is significantly different from simulated mean of means.
Figure 4.F.8: Distribution based on simulated means of change in overall entropy for each condition, including visual comparison between simulated means of means and observed experimental means for Experiment 2 in Chapter 4. Experimental means are significantly different from simulated means of means for all conditions except the majority postnominal condition.

Figure 4.F.9: Distribution of simulated means of change in MI based on head noun across conditions, including visual comparison between simulated mean of means and observed experimental mean for Experiment 2 in Chapter 4. Experimental mean is significantly different from simulated mean of means.
Figure 4.F.10: Distribution of simulated means of change in MI based on head noun for each condition, including visual comparison between simulated means of means and observed experimental means for Experiment 2 in Chapter 4. Experimental means are significantly different from simulated means of means for all conditions.

Figure 4.F.11: Distribution of simulated means of change in MI based on dependent type across conditions, including visual comparison between simulated mean of means and observed experimental mean for Experiment 2 in Chapter 4. Experimental mean is significantly different from simulated mean of means.
Figure 4.F.12: Distribution of simulated means of change in MI based on dependent type for each condition, including visual comparison between simulated means of means and observed experimental means for Experiment 2 in Chapter 4. Experimental means are significantly different from simulated means of means for all conditions.
Chapter 5

General discussion

5.1 Introduction

The overall contributions of this thesis can be summarised in three separate topics. First, by comparing research on cognitive biases across phonology and syntax I established a new division among active cognitive biases into system-wide and category-specific biases. Cognitive biases that affect linguistic structure can be classified into these two types based on the proposed motivations that underpin the existence of those biases, and based on what part of the language they influence. System-wide biases influence people’s preferences for the structural organisation within a language system. For example, system-wide biases in favour of harmony and regularisation/systematicity guide linguistic behaviour to make certain organising principles within a language system more likely to occur and develop over time. On the other hand, there are also category-specific biases that influence our behaviour with regards to specific linguistic items within the wider language system. These category-specific biases are grounded in aspects of semantics and/or processing, such that as a result of these categories conveying specific meanings or exemplifying specific semantic relationships, they have individual biases for how these categories of linguistic items are organised in specific linguistic expressions.

The second contribution of the thesis is the development of the Scale of Innovation. This theoretical construct orders linguistic tasks/contexts along a continuum according to how much innovation is required to perform tasks in that context. Low-innovation contexts include memorisation and pure learning conditions, where the task is essentially to recall an existing linguistic system after training, and high-innovation contexts are those where there is no pre-existing language system in place and each linguistic expression needs to be innovated de
novo. Combined, the division of biases and the ordering of contexts along the Scale of Innovation reveals a tendency for system-wide biases to mainly influence linguistic behaviour in low-innovation contexts, and for category-specific biases to mainly influence linguistic behaviour in high-innovation contexts.

The final contribution relates to the observation that biases which are active at lower ends of the scale have (on the face of it) more opportunities to influence language structure, resulting in population-wide patterns and phenomena. Learning a language system is in many ways a low-innovation context (the system exists, learners receive training/input constructed by their target grammar) and biases which are active at this side of the scale can apply iteratively with each new generation of language learners. The cumulative effects of even weak individual learning biases which apply in this iterative fashion can have substantial effects on population-wide phenomena. In contrast, category-specific biases are mainly observed in the higher regions of the scale, in tasks involving improvisation and language creation. These types of contexts are rare in language, and so they have little ability beyond very early stages of language evolution to actually influence language structure. The fact that category-specific biases can still be observed in cross-linguistic data, but seem to have limited opportunities to shape the linguistic behaviour of language users, creates a missing-link between the biases and their observed effects on typology. This thesis proposed that this missing link can be found in the middle-regions of the Scale of Innovation. In contexts which require innovation, but which also include system learning, we can see some evidence of both types of biases being active at the same time, increasing the number of opportunities that category-specific biases have to influence language structure.

These contributions are informed by the collective results of the experiments in Chapters 2, 3, and 4 of this thesis. In Chapter 2 I identified the existence of a prenominal ordering preference for genitives and a postnominal ordering preference for adjectives. The linguistic context in Experiment 1 was high on the scale, as the suitability of orders were judged by participants in isolation, but did not require them to produce novel structures themselves. In Experiment 2 the same meanings were taught to participants in a gesture-based artificial language learning task with a regularisation design. The main aim was to examine whether participants who were taught languages where the orders matched those that were preferred in Experiment 1 (postnominal adjectives and prenominal genitives) also favoured those same orders in a learning context. This was evaluated by comparing participant’s learning patterns and their tendency to harmonise and regularise the orders they were trained on. The context for the task in Ex-
periment 2, in contrast to Experiment 1, is situated on the left-hand side of the scale, as participants are trained on an existing language structure, although with some variability. The results revealed no effect of category-specific biases on participants’ learning or regularisation behaviour, as participants who learned more bias-congruent languages showed no better learning, regularisation or decreased tendency to harmonise. Instead participants in this task mainly displayed evidence of system-wide biases to increase overall harmony, or reduce unconditioned variation by introducing conditioning contexts. Collectively, the results in Chapter 2 show that both category-specific and system-wide biases are at work for the order of these dependents, but in different contexts.

Since the languages that participants were trained on in Experiment 2 of Chapter 2 were limited to only one of the two dependent meanings (either adjectives or genitives), the experiment in Chapter 3 included a more complex language system where both adjective and genitive meanings were taught to participants during training. The aim of this experiment was to see if the bias for harmony across these two dependent types was affected by the individual category-specific biases for keeping these dependents split across the head noun, meaning that the category-specific biases would be competing with cross-category harmony rather than within-category harmony. This type of competition is potentially what gives rise to the pattern we observe in the typological data for these two dependent types, where the two most common orders are the harmonic postnominal order and the non-harmonic order with prenominal genitives and postnominal adjectives (i.e. the order preferred by the category-specific biases). The results here indicate that we, once again, see little effect of category-specific biases influencing the behaviour of participants during this more complex learning task. Together, these two chapters have targeted parts of the left and right regions of the Scale of Innovation, which left the contexts in the central region unexplored.

This region of the scale is targeted in Chapter 4 in a set of two experiments. The motivation for the experiments in Chapter 4 is grounded in the idea that category-specific biases may have an effect on linguistic behaviour in tasks that combine elements of learning and innovation (regions in the centre of the Scale of Innovation). To explore this possibility a pair of experiments was developed to be directly comparable with each other, while sitting in different regions of the scale. First, Experiment 1 in Chapter 4 was conducted as a conceptual replication of the experiment in Chapter 3, but using only one adjective and one genitive meaning as the basis for the stimuli set. The results of this experiment were broadly the same as for the experiment in Chapter 3 that used the more complex language with two meanings for each dependent type. Following this, Experiment
2, was designed as a generalisation task, which was identical to Experiment 1 in all respects except that participants were tested on the two dependent meanings (one adjective and one genitive) that they did not see during training, paired with two new head nouns. The testing phase in the second experiment thus required within-category extension (the type of task which I have called generalisation throughout this thesis). The results showed that while natural orders were selected at a rate above chance in Experiment 2 and not Experiment 1, there was not a significant difference between the experiments when it came to participants’ learning of natural orders. Nonetheless, the fact that there was a preference for natural orders in Experiment 2 suggests that participants were, to a limited extent, influenced by their category-specific biases for prenominal genitives and postnominal adjectives in the generalisation task. While this preference for natural orders favoured by category-specific biases was more evident in Experiment 2, there were also effects of system-wide biases for harmony and variability reduction in the results of both experiments. Crucially, there was an effect of naturalness on learning when pooling the data for both experiments, such that participants learned the majority order that they were trained on more readily if that order was favoured by category-specific biases. Combined, the results in Chapter 4 show that tasks in the centre of the Scale of Innovation may indeed include influence of both category-specific and system-wide biases, and that learning can (in some instances) be boosted by category-specific biases.

5.2 Further considerations: Methods and assumptions

While this thesis includes experiments that sit along different regions in the Scale of Innovation, it did not target what we may call the “extremes” of either side of this scale. Figure 5.2.1 shows a visualisation of where along the scale the different experiments in this thesis are situated, and it leaves both the right-hand part of the improvising context and the left-hand part of the memorising context empty. This was a conscious choice, and motivated by several different factors. Firstly, preferences in adjective ordering have already been explored in several studies in a production-based silent gesture improvisation context (Culbertson, Schouwstra, & Kirby, 2020; Jaffan et al., 2020), and previous research has shown that the type of judgement-based forced-choice design I employed throughout my thesis has successfully captured similar patterns to those observed in improvisational studies using silent gesture (Do, Kirby, & Goldin-Meadow, 2022; Hughes, 2022;
Therefore, examining the ordering preferences for these dependents using production-based improvisation seemed redundant.

Furthermore, a significant amount of the experimental work in this thesis was also conducted during the COVID pandemic. Performing gestural improvisation experiments would have involved recording people in their homes using their own camera equipment and set-up. While the technology to conduct automated experiments of this kind is now available, the approximation I employed using forced-choice testing with gestural stimuli allowed for simpler ethics, faster data collection, and a higher number of participants per condition due to the reduced cost and potential for data loss. Additionally, forced-choice tasks can be seen as a conservative measure of the kinds of cognitive biases that are observed in improvisation experiments. During production-based tasks participants are not given an option between those structures which are proposed to be favoured and disfavoured by the bias(es) under investigation. Participants may therefore not entertain the possibility of the unnatural/disfavoured variant during their productions, meaning you would expect to get stronger category-specific effects in production-based improvisation experiments where both options are not presented to the participants. The fact that the forced-choice methodology still captured effects of category-specific biases when both variants were presented to participants can thus be seen as even stronger evidence for the existence of these ordering preferences. Using a production-based task to examine the interplay between these two types of cognitive biases would therefore predict a more extensive preference for the category-specific biases in the first experiment of Chapter 2 than what we see using a judgement-based task. Additionally, considering the ongoing debate regarding the source of regularisation effects (recall outline of this in Chapter 1 section 1.5.4), the use of a production-based task in a learning experiment like the one in Chapter 3 could help elucidate if regularisation effects are driven by the mechanism of learning or production. If regularisation behaviour is mainly motivated by production constraints, then we would expect to see greater use of regularisation strategies in a production-based versions of this experiment. If, on the other hand, there is no difference in levels of regularisation between production-based testing and forced-choice testing, then production constraints would not be considered the driving force behind regularisation effects.

Considering the memorising region, which sits at the other extreme of the scale, tasks in this region involve “pure” learning, and potentially wholesale memorisation of a language system. As such, experiments that focus on this region tend to examine ease of learning between language systems with different prop-
erties, to evaluate if some of those properties impart a learning advantage. These different properties tend to be consistent across the system, i.e. the comparison between conditions is between languages that either have some property, or they do not. While this type of methodology could have been used to examine the learnability of bias-congruent and incongruent word orders for adjectives and genitives, the likelihood that this would have revealed interesting differences is slim. Ease of learning experiments have a tendency to fail due to ceiling effects, whereby participants across the conditions learn the language to such a high proficiency that potential differences are masked (e.g. as was observed for adults in the study by Tal and Arnon (2022) which examined potential learning advantages for redundancy in language). Given that learning differences are already minor between the conditions in the regularisation experiments in this thesis, paired with the fact that I was interested in participants’ harmonisation and regularisation tendencies in addition to potential learning advantages afforded by category-specific biases, ease of learning was not a suitable paradigm to examine the type of competing biases that I focused on in this thesis.

When designing the experiments in this thesis I essentially “skipped” an entire region of the scale by going from an experiment in the improvising region, to an experiment in the generalising region. A generalisation task is in many ways more of a hard test for the influence of category-specific biases, since it is lower down on the Scale of Innovation. Again, the fact that we still see some effects of these biases in that context is testament to their ability to influence linguistic behaviour. However, this result also necessarily implies that we should see influence of these same biases in an extrapolation task. This task could, for example, include training on only one dependent type (as in Experiment 2 in Chapter 2) but test participants on the order of nouns and dependents for the other dependent type. Given the increased meaning distance from members of the same category to members of different categories that this type of task would involve, we would expect even more influence from category-specific biases in that kind of context (recall mention of this in the Discussion section of Chapter 4). Future examinations of the effects of category-specific biases in this type of context would be highly beneficial to solidify the motivation for the ordering of contexts along the Scale of Innovation.

A key difference between the first experiment in this thesis, where I tested the existence of the category-specific biases for genitives and adjectives, and all subsequent experiments, is that the original experiment was a single-measure task. Participants in this experiment only had to select a gestural expression for one meaning, whereas all the other experiments involved repeated selections for sev-
eral different meanings. The lack of influence of category-specific biases in several of the regularisation experiments tasks could therefore be argued to reflect a difference in measure, rather than a true difference in activation of category-specific and system-wide biases caused by the change in task-based innovation. It is possible that category-specific biases are especially strong during isolated linguistic events, since repeated measure tasks may initiate the building of a linguistic system, causing system-wide biases to begin influencing participants’ behaviour. However, two things contradict this assumption. Firstly, there is still some influence of category-specific biases in the generalisation task in Experiment 2 in Chapter 4, although this was not significantly different between the regularisation and generalisation experiments. Crucially this experiment does involve repeated measures for the test items. Furthermore, I conducted an exploratory analysis where I extracted the first test trial for each participant from the regularisation experiments across my whole thesis. With this data I generated a binary outcome variable (\( \text{natSelected} \)) where 1 was prenominal order for genitives and postnominal order for adjectives, and 0 was the reverse of this. The model also included dependent type as a two-level fixed effect and experiment a four-level fixed effect (both were deviation coded) as well as their interaction. Model comparison using a likelihood ratio test revealed that the null model (intercept only) was the best fit for the data, and that including dependent type, experiment or their interaction did not improve model fit (\( \chi^2 = 2.012 \ p = 0.16; \chi^2 = 7.28 \ p = 0.06; \chi^2 = 2.25 \ p = 0.52 \)). The intercept only model was not significant (\( \beta = 0.04, \ SE = 0.07, \ z = 0.57, \ p = 0.57 \)), showing that participants across all the regularisation experiments did not select more natural orders in their first trial than is predicted by chance. This suggests that the preference for natural orders from the first experiment in Chapter 2 is not just caused by the difference between a one-shot measure and a multiple measure task. Other differences between the experiments,
such as the system training that all participants in the regularisation experiments had, may contribute to the observed difference in naturalness effects across the different experiments in this thesis.

Considering the implication of task choices for the results of this thesis is important, but the same is also true for the choice of meanings that the artificial language denoted in each experiment. All the experiments in this thesis specifically targeted adjective meanings and genitive meanings, since word orders with these two dependent types include interesting typological asymmetries that suggest they are influenced by both category-specific and system-wide biases. However, two questions may arise when considering the status of these dependents across the experiments in this thesis. First, do the stimuli employed in the experimental tasks activate the intended meanings usually ascribed to these categories (descriptive meanings for adjectives and possessive meanings for genitives)? Secondly, are these meanings represented by the categories “adjective” and “genitive” in our participants’ minds? The first question was one I worked hard to alleviate during the piloting stage for these experiments. This is one of the main reasons that we included the translation trials in the early experiments, even though this data was not intended for formal analysis. In these trials participants were shown a gesture video signifying a specific meaning and asked to translate the meaning into English. In some of the experiments they were only shown adjective or genitive meanings and in some instances they saw both. By monitoring participants’ responses to these translation tasks, I was able to fine-tune the experimental instructions and study design to maximise the likelihood that participants would give descriptive (adjective) and possessive (genitive) translations for these trials. This suggests that the tasks in the experiments successfully activated the intended meanings for many participants. Indeed, showing people individual images of the kinds of relations we were interested in did not seem to activate the intended meanings. When participants saw just an image of a vampire character wearing a hat they did not necessarily infer a possessive relation between the vampire and this object, but rather tended to produce active relations such as “vampire wearing a hat”. The piloting is what eventually led to the use of a combination of grid-based contrast trials at the start of the experiment to give participants an opportunity to study the difference in relations between stimuli items, and also including this grid during critical trials. It is worth noting that this did not eliminate participants’ tendency to give verb-like translations for the genitive relations, or adpositional phrases for adjective meanings, but it did make the intended meanings more likely. Furthermore, although the translation trials were intended as a check for how participants interpreted
these relations it is important to note that as it forced them to use English to express these relations, this act of translation might have altered the way they truly represented these meanings when only interacting with the gestural stimuli.

The categorisation of the translation data that is available in Appendix A, section 2.E and Appendix B section 3.F suggests that participants expressed the meaning of gesture videos in a number of ways, and that these translations often did not directly parallel the categories adjective and genitive. However there are several reasons for why these translations should not be taken as clear indications that participants represented these meanings as the wrong categories. First of all, while the number of direct genitive translations for possessive meanings is low, there was a high number of structures which fell into the “Other” and “VP” categories and that still expressed a possessive meaning. These include possessive VPs like “vampire has a hat” and plain juxtaposition “vampire hat,” which is a structure used in many signed languages to express possessive relationships (Perniss & Zeshan, 2008). These structures may very well have been represented as genitives in the minds of participants, but have been transformed into these other grammatical structures by the act of translation. Furthermore, it is possible that, rather than integrating these meanings into existing grammatical categories, participants instead see the gestural representations in a way that is similar to parataxis, where spotted cup is really represented as something more like “there is a cup” and “the cup has spots”. This paratactic interpretation has even been suggested to constitute an evolutionary stage during the emergence of new signed languages which then gradually develops to become fully grammaticalised syntactic forms in a mature sign language (Hall, 2017; Sandler, 2017). As such, even if participants are not assimilating these gestural expressions with the exact categories of adjective and genitive that exist in their own native first language, their interactions with these meanings still speak to a meaningful interpretation of linguistic content that has naturalistic equivalents in the evolution of signed languages.

Crucially, these translations were not prompted by the same stimuli as the meanings in the experiments themselves. In the main testing phases people were associating images with adjective and genitive meanings, whereas the translation trials used the gesture videos to prompt participants’ productions. As such, the translations can tell us little of how the meanings conveyed by the images were encoded by participants. In future research, including both the target image and the gestural expression of that meaning as prompts in the translation trials would help mitigate this issue. As it stands, a note of caution is needed when interpreting the present results as true reflections of the way participants convey
genitives in particular. It is possible that the observed prenominal preference is a partial reflection of the subject-first bias (Futrell et al., 2015; Meir et al., 2017), if participants are interpreting these meanings as verb-like, since this would also place the possessor first in the structure.

Some of the effects we are seeing in the experiments in the thesis could be specific to the modality of the stimuli used in the experiments. For example, the previously discussed preference for SVO/SOV order for expressing basic transitive events has mainly been studied experimentally using gestural studies. More recent work suggests that this preference does not necessarily carry over into the vocal/auditory modality, or at least it is not as strong (Struhl et al., 2017). This difference between the modalities is not unexpected given the connection between the animate/human first preference and the tendency for gestures denoting humans, or other highly animate referents, to make use of the body of the gesturer/signer as the individual being denoted (Meir et al., 2007, 2010, 2017). This adds an element of “me first” on top of the existence of the general animacy linearisation preference. In the studies reported in this thesis participants are not themselves producing the gestures, so the preference for possessors to precede the head nouns is not linked in the same direct way to a motor-planning preference for starting from the participants’ own body. However, it is possible that the strengthening of this preference may still exist for gestural stimuli, even when participants themselves are not producing the gestures.

Similarly, the adjective meanings used in this thesis were limited to those denoting patterns (spotted and striped). Previous work using production-based gestural studies has found that these adjectives tend to exhibit an especially strong postnominal preference (Culbertson, Schouwstra, & Kirby, 2020). This is mainly an issue when the nouns used in combination with these adjectives lend themselves to being drawn in space (e.g. a shape like a rectangle or circle) since participants tend to draw the noun and then “fill in” the shape with the pattern. The experiments in this thesis use objects as nouns rather than shapes, which makes the “fill-in” strategy less available and adjectives were not conveyed “inside” the shape of the nouns in this way in the gesture videos. These design choices mean that the postnominal adjective preference observed in the thesis is not simply the result of a modality effect where gesture allows for this type of “fill in” strategy for pattern adjectives, but rather reflects a more fundamental preference for postnominal order.

Experimental work often suffers from issues regarding the true generalisability of their results, given the specificity of the phenomena it studies and the population it tests. In my own research, I had to balance the availability of my testing
population against the generalisability of my results. While all my participants report English as their first language, I placed no restriction on other spoken languages that they might know. I chose this population as I had easy access to a large number of participants that fit these criteria online, a necessity as the majority of the data collection for my thesis was conducted online during the COVID-19 pandemic. In the future, similar to how the research on the biases for harmony and homomorphism now includes speakers of a more diverse set of languages (Culbertson, Franck, et al., 2020; Martin et al., 2019, 2024), I hope to study similar phenomena in more diverse populations.

With that said, English speakers are actually a fairly good population to test for the specific biases I focused on in this thesis. The canonical placement of adjectives in English is prenominal, meaning that the native language biases is the opposite of the proposed cognitive bias for this dependent. With regards to genitives, English has variable word order for possessive constructions where the ‘s possessive is prenominal and the of possessive is postnominal. This means that, although the proposed favoured order for genitives (prenominal) is available in participants’ native language, it is not the only possible order. An ideal population to test for the category-specific biases would be one where the only available orders are prenominal adjectives and postnominal genitives. A possible target population with this language structure would be speakers of Polish. Genitives follow the head noun in Polish noun phrases. For example, in the phrase książka dziewczyny meaning ‘the girl’s book’, the possessed item (the head noun) is the initial word, followed by the possessor. Adjectives, on the other hand, tend to precede the head nouns in noun phrases. For example, in the phrase czerwony kapelusz meaning ‘the red hat’, the colour adjective is the initial word, followed by the head noun meaning ‘hat’. While other orders are available under some circumstances (just like in English) the fact that the default orders are the opposite of the proposed category-specific biases makes Polish speakers a good population to use as the hardest test for the existence of these biases two category-specific ordering biases.\footnote{Note that the most natural translation for a phrase like ‘spotted cup’ in Polish would be to describe this using an adpositional phrase, \textit{kubek w kropki}, where the adjective does follow the noun. Using a different set of adjectives to those used in this thesis may be needed to ensure that the default positioning in participants’ L1 is the opposite order of the one favoured by category-specific biases.}

It is worth noting that the majority of the word order data that I have presented in this thesis is based on counts of languages that have been assigned a categorical word order feature (e.g. either prenominal or postnominal/ OV or VO). This method comes with certain drawbacks, such as the fact that most lan-
languages have some degree of flexibility in word order which is not captured by a binary categorisation. Levshina et al. (2023) recently produced a detailed discussion of how gradiency should be the default assumption for word order analyses across languages, and calls for a more nuanced treatment of these tendencies than the current data provides. They highlight that languages like English and Indonesian are comparatively consistent in the order between elements like Subject and Object (both languages have nearly 100% Subject-Object order) whereas other languages like Lithuanian and Tamil have much more variability in the order of these elements (Lithuanian having just over 60% Subject-Object order, and Tamil having just over 70% Subject-Object order). Including this consideration of gradiency in our studies is potentially highly beneficial as it follows other trends in the field. Like the change in the field from discussions of absolute universals to typological tendencies reflects the probabilistic nature of cognitive biases, the probabilistic impact of cognitive biases is also reflected within features for each individual language. However, currently there is no comprehensive typological database which allows for a gradient measure of word order to be used as a basis for hypothesis generation of the kind I conducted for this thesis.

In addition to the potential issue related to categorical features, the count data presented in this thesis does not take into consideration the deep historical relations between languages, meaning that the samples are potentially biased by shared ancestry. Previous research has shown that several typological patterns are (at least partially) correlated with phylogenetic history (Dunn et al., 2011; Hartung et al., 2022). This is the precise reason that combined data from experimental work and typological data is so important; it is only in the combination of these sources of evidence that we can establish if there truly is a behavioural bias involved in the patterns we see in typology. By just looking at the distribution of word orders for adjectives and genitives in relation to nouns all I could say is that it looks as though there are competing influences, but only with the additional experimental data showing that there are individual biases for these dependents can we claim that the typological pattern may have cognitive roots. Furthermore, phylogenetic and cognitive explanations of typological patterns are not mutually exclusive. Features that emerged because they are favoured by cognitive biases may also be more likely to survive over time, and if those proto languages went on to diverge into many modern languages, then those favoured features will also be correlated with language history.
5.3 Future directions

Future application of the suggested division of cognitive biases into system-wide and category-specific biases should answer the following three questions when establishing which type of bias a given behaviour/phenomenon can be classified as:

i What grounds/motivates the bias?

ii What is the bias’ domain of influence?

iii Under what conditions does the bias influence behaviour?

These questions form interdependent answers. If a bias is motivated by principles outwith its domain of influence it will, in all likelihood, be a category-specific bias. The examples from this thesis include the semantic motivation behind homomorphism, the processing motivation behind the postnominal adjective preference and the phonetic/articulatory motivations behind certain phonological regularities. On the other hand, if the bias is motivated by factors internal to its domain of influence it will, in all likelihood, be a system-wide bias. Examples from this thesis include the motivation for syntactic simplicity grounding the bias for word order harmony, and the motivation for phonological simplicity grounding the bias in favour of phonological alternations with few active features. The domain of influence also differs between the two bias types, such that system-wide biases result in preferences for specific systematic organising principles that span across a language system (such as feature distribution), whereas category-specific biases influence the structure and expression of individual linguistic items, rather than principles that extend across different categories of items in a language system. Finally, the different biases are also predicted to be stronger/more active in different contexts. Category-specific biases are mostly active in high-innovation contexts where the absence of a wider linguistic system, or direct evidence for how to produce a structure, makes people fall back on the grounding principles of these biases (semantics, processing, ease of articulation etc.). System-wide biases, on the other hand, are especially active during tasks that involve the acquisition and consideration of a wider linguistic system, where the benefits of the features favoured by system-wide biases are most beneficial as organising principles (harmony, feature distribution etc.).

Categorising biases in this way allows for predictions to be made about where along the Scale of Innovation you should be able to observe the bias having an effect. This allows for both a better understanding of the biases’ potential
role during language evolution, and helps future researchers select appropriate experimental methods to study the phenomena they are interested in.

While the Scale of Innovation may act as a useful methodological tool for researchers who are interested in studying different types of cognitive biases, I also want to spend some more time elaborating on the idea that there may be a more direct parallel between the regions along the scale and proposed stages of language evolution. This parallel would make the scale especially interesting to researchers who are interested in studying how cognitive biases have played a role in shaping language structure over evolutionary time. While the Scale of Innovation moves from left to right, in the sense that the amount of innovation increases as you move from contexts on the left to contexts on the right, this can be matched by a reversed scale of the progression of language evolution which starts on the right and progresses to the left. Figure 5.3.1 shows an expanded version of Figure 1.3 from Chapter 1, which now includes this matched scale displaying the progression of language evolution.

Figure 5.3.1: Scale of Innovation mapped to proposed progression of language evolution. The top scale progresses from left to right, as innovation increases, and the bottom scale progresses from right to left as the evolution of language moves from the very first stages of language creation to the fully established conventionalised systems we see in modern languages.

Contexts on the far right of the scale are those where the need to innovate is highest. In these contexts there is no preexisting language system, and all linguistic expressions must be innovated de novo. This parallels the very earliest stages of language evolution, where there is no conventionalised language, and people simply had communicative goals and used the articulatory tools available to them to convey their intended meanings. Many who argue for a gradual, rather than catastrophic, development of linguistic capabilities tend to support

\[ ^{2}\text{Some argue that language evolved in a catastrophic way, i.e. the mechanism which allowed for the evolution of language to occur happened by sudden mutation in one individual (Berwick, 1998; Chomsky, 2005).} \]
the notion that the original “proto” language(s) went through stages of increasing structural complexity, starting with a one-word stage, and then developing to two-word stage and eventually short phrase stage before expanding to a full sentence stage (Hurford, 2012; Jackendoff, 2002; Pinker & Bloom, 1990). In this way, lexical items are thought to have been established much earlier during language evolution than subsequent syntactic structure, an argument which is exemplified by Givón’s statement that language would have had a “lexicon before grammar” (2002, p. 157). This gradual development of lexical items would have lent itself to the next stage, namely extrapolating, which involves the innovation of novel categories of items and/or combinations of items to cover new meanings that have not yet been systematised. Similarly, following this stage we may have seen an enrichment of these established categories during a period of generalisation. This then leads on to a period where there is an established language, but where much individual variability is present. This variability then needs to be negotiated and regularised across a wider population of language users. Finally, once a fairly stable form of the language has been conventionalised within a community, this system can then be passed on faithfully to new language learners, although it will still undergo normal processes of language change. This final stage is the stable yet dynamic state that modern languages are now in and relies heavily on system learning. Crucially, the last region called memorising on the scale is an idealised version of this final stage, since language learners cannot fully memorise the language system(s) they are acquiring.

While the Scale of Innovation can be loosely mapped to this idea of language evolving lexical and syntactic complexity over time, (starting from a one-word lexical stage and developing syntactic structure gradually) this view has been challenged by those who consider a different notion of language origins (Tomasello, 2003). According to this view, language evolved in a holistic way, starting from a point where individual “words” and linguistic signals conveyed complex semantic meanings (see e.g. Wray, 1998). These signals would have been non-compositional, as sub-parts of the signal could not be associated with sub-parts of the meaning (Smith, 2008; Tallerman, 2007). This theory is supported by some modelling and experimental work on iterated learning and cultural transmission, since these studies find that a holistic starting language can evolve to have compositional structure over generations in both artificial agents (see Kirby et al., 2004, for an earlier review) and human participants through the process of cultural transmission (Kirby et al., 2008, 2015). Further support is found in studies of how children acquire language by memorising larger sequences, or “chunks” of language, which they then gradually analyse and come to realise are made up
of distinct parts that each correspond to different parts of the meaning (Arnon & Snider, 2010; Arnon et al., 2017). This view of acquisition as “starting big” (Arnon, 2021) draws on construction-based ideas of language structure (see e.g. Goldberg, 2006), and can be seen as an ontological parallel to the progression of language evolution from a holistic perspective.

Although evolutionary theories that support this holistic view of the origin of language are crucially different from the lexical view I outlined in the earlier paragraphs above, there are still interesting parallels between the holistic view and the Scale of Innovation. For example, the initial stage of language evolution maps well to the improvising contexts independently if the meanings that individuals convey in that context are complex, and whether the utterances are holistic or not. The extrapolation and generalisation stages of the scale may instead relate to what parts of the linguistic signals and complex meanings are first analysed and made compositional. The key question is if this holistic starting point would prevent some category-specific biases from influencing early language evolution, since their activation may require the development of compositionality to have progressed to a point where some notion of different linguistic categories are in place. Examining if the order of analysis (i.e. the decomposition of holistic representations) is affected by category-specific biases to varying degrees based on how much of the system has already been analysed could present an interesting avenue for future research that wishes to apply ideas from this thesis to test theories with a holistic view of language evolution and a “starting big” approach to language acquisition.

In my outline of the Scale of Innovation I purposefully labelled areas within that scale to align with different experimental paradigms. This was done to highlight the parallel between some of the proposed evolutionary contexts that resemble these experimental methodologies, thereby showing how using these methods help shed light on processes in language evolution. Like with any continuum there are prototypical instances of contexts within each of the labelled regions on the scale, but there are also borderline cases or examples of mixed tasks. These borderline tasks involve a combination of elements from different parts of the scale. Previous studies have already made use of this type of combined methodologies (see e.g. Do, Kirby, & Goldin-Meadow, 2022; Motamedi, Wolters, Naegeli, et al., 2022), and Chapter 4 Experiment 2 is an example since it used a regularisation design in the training data, but a generalisation task at test, combining features from both of these regions of the scale. Future studies could continue to explore these combination tasks. For example, one could imagine an experimental task where participants are trained on all the lexical items in a system, but never see
how any of them are combined to form more complex phrases. In this instance, all
the building blocks of the language system are available to participants (as in the
learning context) but none of the systematic structure of the language (as in the
improvisation context). This type of experiment is similar to some of the extrap-
olation studies I have already mentioned (e.g. Culbertson & Adger, 2014; Martin
et al., 2020; Saldana et al., 2021) but differ in that the order between none of
the elements would be included in training. Exploring this type of method could
further improve our understanding of how this scale operates, and the connection
between different features that define contexts as belonging to different regions
along the scale. There has already been some research employing this type of
method (Marno et al., 2015; Motamedi, Wolters, Schouwstra, & Kirby, 2022)
using gestural stimuli, and in these contexts they saw a rise in the influence of
participants’ native language influence on the structures they produced at test.
It is possible that this isolated lexical learning means that, instead of starting to
formulate a new grammatical system, participants simply map the new linguistic
signals onto pre-existing words/signs in their native language(s).

Another source of evidence which could help validate the idea that category-
specific biases are especially strong in high-innovation contexts would be to per-
form an in-depth study of the patterns of genitive and noun ordering and adjecti-
tive and noun ordering in young signed languages. As mentioned previously, these
contexts are the closest modern equivalent we have to newly evolved languages,
and are therefore an invaluable source of evidence for researchers interested in
language evolution. If we were to examine the two category-specific biases I have
focused on in this thesis in different cohorts of speakers of Nicaraguan Sign Lan-
guage (NSL) we would expect to see mainly prenominal genitives and postnominal
adjectives in early cohorts. Tracking potential changes to noun phrase word order
across cohorts would help inform us about the changing pressure from different
biases as more and more of the language system became conventionalised within
the language population. Some initial results from research investigating adjec-
tive ordering suggests that there is a postnominal adjective preference in NSL,
but this work did not include an explicit investigation of genitives (Do, Kirby,
Goldin-Meadow, et al., 2022). Other previously identified category-specific pref-
ences for event-type conditioned word order (SVO for intensional events and
SOV for extensional events; see (Schouwstra & de Swart, 2014)) have been ex-
amined in signed languages with similar goals in mind. The same event-type
conditioned pattern has been observed in some young signed languages, such as
Nicaraguan Sign Language (Flaherty et al., 2018), and even in some older signed
languages like Brazilian Sign Language (Napoli et al., 2017). In contrast to the
category-specific biases I examined in this thesis, the event-type conditioned word order pattern is not clearly observable in the typology of spoken languages. It is possible that the system-wide bias that event-type conditioning competes with is stronger than the system-wide bias for harmony. Motamedi, Wolters, Naegeli, et al. (2022) highlights this possibility and argues that the pressure to converge on a shared basic word order may be especially strong due to the communicative benefits that consistency in word order brings in the absence of clear morphological marking. However, if we consider the arguments of Levshina et al. (2023) it is possible that if variability in word order was examined in more depth in spoken language, that there may be more residue of event-type conditioning than categorical word order categorisations imply.

Similarly, the same type of investigation could also be conducted on homesign systems. Recall, homesign systems emerge when Deaf children, who are linguistically isolated, innovate their own manual language systems to communicate with their caregivers/family, and they do this without real access to any other language model. As in the case of young signed languages, we would expect more effects of category-specific biases in the earliest expressions of homesign systems, such that they show strong biases for using transparent/iconic representations of the meanings they try to convey, and that their productions would be strongly influenced by attention and processing considerations. With regards to the work in this thesis, we would therefore expect to see a strong preference for prenominal genitives and postnominal adjectives across homesign systems. Additionally, studying homesign would allow for a more precise investigation of what it means for system-wide biases to have an effect on language. Since homesign systems appear to be the product of a single individual creating a language (Carrigan & Coppola, 2012; Goldin-Meadow et al., 1984), it would be possible to track the order in which they create expressions for different meanings, and how much or which parts of that system need to be in place for system-wide preferences to influence the structural organisation of the language. As it stands, it is unclear how large, or how connected, a language system needs to be in order for system-wide biases to have an effect on behaviour.

In addition to this, the early productions by children using homesign could also be a powerful set of complementing evidence to the data from young sign languages. Comparing the effects of category-specific and system-wide biases in these two contexts would help us understand the role that reciprocal linguistic events, and community-level alignment, play in revealing the effects of both category-specific and system-wide biases on language structure. For example, the influence of the preference for harmony may be especially strong in linguistic
contexts that involve communicative events between multiple speakers/learners, since establishing a joint system or ordering rules helps promote successful communication between these different individuals. This need for alignment is not necessarily as important in homesign systems with only one proficient user and a limited number of interlocutors. The idea that there would be a difference in the level of variability that a young signed language and a homesign system can sustain is supported by research which shows that languages with a smaller set of users, and more tightly connected social networks, can sustain more individual-level variation compared to languages with more users and more loosely connected social networks (Meir et al., 2012; Raviv et al., 2019). Examining what these differences between homesigns and young signed languages means for the effects of system-wide biases on language structure would be an interesting way to expand the study of these cognitive biases from learning and innovation contexts, to contexts involving different types of communication tasks and social structure.

In general, this thesis has focused on the tension between cognitive biases which are active during tasks that rely heavily on language learning, and tasks that rely on heavily on innovation. I was particularly interested in this combination of biases due to their differing opportunities to influence typological structure. However, as I noted in Chapter 1 and in the paragraphs above, learning biases are not the only type of bias that has extensive and iterative opportunities to influence language structure. Biases which influence our behaviour during language use, and communication also have the ability to have strong structure-shaping effects on language (Fedzechkina et al., 2023; Ferdinand et al., 2019; Gibson et al., 2019; Hawkins, 1983; MacDonald, 2013). These are important aspects of cognition that also require further study, and while the Scale of Innovation was developed to focus on learning and innovation contexts, similar reasoning could potentially be applied to some communicative pressures active during language use. For example, different communicative contexts may require more or less social innovation in terms of how much common ground or shared knowledge that interlocutors have in that specific context. Similarly, inferences about our interlocutor’s language knowledge may make us more or less likely to employ specific communicative strategies. The possibility that biases which are active during communication may also be modulated by the amount of innovation involved in the context (dictated by the specific communicative event and specific interlocutors) requires further investigation.

In addition to this extension of the applicability of the scale of Innovation to situations involving communication, there are also parallels between the tension represented by learning and innovation, and work examining how other types of
trade-offs shape language structure. Extensive work has focused on how language tends to self-organise along an optimal frontier (often referred to as the pareto-frontier) between two competing forces, for example simplicity (sometimes called compression) of the language system and expressivity (sometimes called informativeness) of the meanings the language can encode and convey (Kirby & Tamariz, 2022). The result of this tension has been useful in explaining a number of features of language, such as the emergence of combinatorial structure (Kirby & Tamariz, 2022; Kirby et al., 2015), the shape and content of colour and kinship systems across typology (Gyevnar et al., 2022; Kemp & Regier, 2012; Kemp et al., 2018; Zaslavsky et al., 2018), how person spaces and grammatical markers are encoded in different languages (Mollica et al., 2021; Zaslavsky et al., 2021), and the possible meanings of scalar adjectives (Carcassi et al., 2019). If we consider satisfaction of category-specific biases to be the dimension of expressivity, and satisfaction of system-wide biases to be the dimension of simplicity, then the parallels between the trade-off found in the Scale of Innovation and the rest of the literature above seems quite straightforward. However, there is a key difference in how these pressures are operationalised which differentiates the tension between system-wide and category-specific biases that I find in my work, and the tension between simplicity and expressivity that has been discussed previously. Crucially, satisfaction of category-specific biases does not necessitate an increase in system complexity in the way that the trade-off literature above tends to argue. For example, languages that mark more colour distinctions have more complex colour naming systems, i.e. to satisfy the need for expressivity languages have to sacrifice along the dimension of simplicity (Gyevnar et al., 2022; Zaslavsky et al., 2018). In contrast, as I showed in section 1.6 with regards to homomorphism in the noun phrase, the category-specific bias for transparent linearisation of semantic composition can be perfectly satisfied while also satisfying the system-wide bias for harmony. This is exactly what we see in the two most common word orders across languages, namely Noun-Adjective-Numeral-Demonstrative and Demonstrative-Numeral-Adjective-Noun. These two orders are both harmonic and homomorphic, and so satisfy both biases without sacrifice along either dimension of expressivity or simplicity. Despite this difference, the way that information bottlenecks (limits to the amount of the system that is passed on at each generation) are used in this research means that there is potential to expand the parallel between the work above and the concept of the Scale of Innovation. By manipulating what part of the system that the information bottleneck targets, this research could compare how, for example, generalisation vs extrapolation contexts influence the trade-off between expressivity and simplicity.
As outlined in section 1.4 of Chapter 1, both system-wide and category-specific biases have been independently examined across syntax and phonology. Although the grounding and quantification of these biases differ across the two domains, the research goals and questions share striking similarities that, if syntactic and phonological work were to be combined, would make for an interesting and cohesive view of how cognitive biases influence several domains of language structure. Due to this, an interesting future direction would be to design a set of phonological experiments that also map to different regions in the Scale of Innovation. Ideally, a similar instance of competition between a system-wide bias and a category-specific bias would be identified to make the research maximally parallel to the work in this thesis. This type of design would allow us to examine if the cut-off point for where we see which type of bias affecting behaviour is the same across phonology and syntax. Such evidence would help us evaluate how generally applicable the Scale of Innovation is for studies of cognitive biases across linguistic domains, and if changes to the level of innovation in a task is comparable across those same domains. Finding similar results in the phonological domain would establish the scale as a useful design concept for future researchers interested in both cognitive biases and language evolution. Depending on the grounding of the suggested bias, researchers can choose experimental tasks which are more likely to reveal the behavioural effect of the bias they are interested in. Similarly, researchers who are interested in a specific period of language evolution can base their experimental design on how far along the scale the evolutionary period lies.

5.4 Conclusion

This thesis has introduced and described a two-tier classification system for cognitive biases that influence linguistic structure. System-wide biases are motivated by factors internal to the domain in which they have an effect (e.g. syntactic motivations for syntactic effects) and they influence the organisational structure across a language system. Category-specific biases are motivated by factors external to the domain in which they have an effect (e.g. phonetic motivations for phonological effects) and they influence the expression of individual linguistic items or meanings. By dividing biases in this way we can make predictions regarding the types of contexts in which we might expect to see the different biases influencing linguistic behaviour. To formalise the consideration of different contexts I proposed the existence of the Scale of Innovation, a systematic ordering of language contexts and experimental tasks based on the amount of
innovation that is required to complete linguistic tasks in those contexts. Based on previous observations from experimental literature on cognitive biases I proposed that system-wide biases should mainly influence individuals’ behaviour in low-innovation contexts, whereas category-specific biases should mainly influence individuals’ behaviour in high-innovation contexts. Furthermore, I highlighted that in order for category-specific biases to be reflected in present-day typology their influence cannot be limited to improvisation contexts, since these effects are predicted to have been washed out by the continuous application of competing system-wide biases over time. This led to an investigation of the possibility that category-specific and system-wide biases may both be active in contexts in the middle region of the Scale of Innovation that require both elements of learning and innovation.

Using a case study, informed by typological data on noun phrase word order, I investigated how a system-wide bias for harmony competes and interacts with two category-specific biases influencing ordering preferences for two noun phrase dependents, namely adjectives and genitives. I found evidence for the influence of the bias for harmony in participants’ behaviour during regularisation tasks of varying complexity (low-innovation contexts), but little to no influence of category-specific biases in these same tasks. In contexts requiring more innovation, such as expression judgement tasks without a surrounding language system (high-innovation context), participants were guided by the category-specific ordering preferences for both postnominal adjective order and prenominal genitive order. Interestingly, in contexts in the centre of the scale that combine elements of learning and innovation, the harmonic bias persisted but also seemed to compete with the effects of the category-specific biases, which favoured keeping adjectives and genitives split across the head noun. This last set of results provides the first indications of the missing link between the contexts in which we see category-specific biases influencing behaviour and their ability to influence typological regularities. Contexts that combine learning and innovation occur more often in natural language contexts than situations of language creation, meaning that both category-specific and system-wide biases have recurring opportunities to influence language structure.

This thesis thus provides a new theoretical tool, the Scale of Innovation, that researchers can use to examine the effects of different types of cognitive biases, both within individuals and when considering population-wide phenomena in linguistic typology and language evolution. The division between system-wide and category-specific biases also captures the duality of language as being both a complex, rule-governed system and a continuously changing and creative expres-
sion of human thoughts and intentions that places different demands on language users in different contexts. The case study used in this thesis acts as a starting point for employing this framework in future studies of cognitive biases across different linguistic domains.
Thesis bibliography

NOTE: The following bibliography encompasses all works cited within this thesis. Due to this, any citations with the same author/year which have a letter next to the year (e.g. Dryer, 2013a-e) will not necessarily match up with the letters assigned in each individual chapter. Any individual examination of a source based on either of the main content chapters should be directed to the sub-bibliography in the associated chapter. The chapter bibliographies include only the sources cited within that specific chapter.
Bibliography


279


281


Culbertson, J., Franck, J., Braquet, G., Barrera Navarro, M., & Arnon, I. (2020). A learning bias for word order harmony: Evidence from speakers of non-


Dryer, M. S. (2013a). Order of adjective and noun. In M. S. Dryer & M. Haspelmath (Eds.), The world atlas of language structures online. Max Planck Institute for Evolutionary Anthropology.

Dryer, M. S. (2013b). Order of adposition and noun phrase. In M. S. Dryer & M. Haspelmath (Eds.), The world atlas of language structures online. Max Planck Institute for Evolutionary Anthropology.

Dryer, M. S. (2013c). Order of genitive and noun. In M. S. Dryer & M. Haspelmath (Eds.), The world atlas of language structures online. Max Planck Institute for Evolutionary Anthropology.

Dryer, M. S. (2013d). Order of numeral and noun. In M. S. Dryer & M. Haspelmath (Eds.), The world atlas of language structures online. Max Planck Institute for Evolutionary Anthropology.

Dryer, M. S. (2013e). Order of subject, object and verb. In M. S. Dryer & M. Haspelmath (Eds.), The world atlas of language structures online. Max Planck Institute for Evolutionary Anthropology.


285


Durvasula, K., & Liter, A. (2020). There is a simplicity bias when generalizing from ambiguous data. *Phonology, 37*(2), 177–213. https://doi.org/10.1017/S0952675720000093


286


Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and


293


298


299


302


Wang, F., Kirby, S., & Culbertson, J. (2023). The learning bias for cross-category harmony is sensitive to semantic similarity: Evidence from artificial lan-


308


