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.
From cognition to word order universals: An artificial language learning approach

Fang Wang

Doctor of Philosophy
School of Philosophy, Psychology, and Language Sciences
University of Edinburgh
2024
Abstract

Despite their striking diversity, human languages exhibit many statistical universals. Explanations for the origin of statistical universals include language relatedness, innate properties particular to the language faculty of the human mind, general cognitive biases, functional pressures, and diachronic processes. Evidence for these different explanations mainly comes from historical and typological data. However, causal links between specific hypothesized mechanisms and particular language features are difficult to establish based on these data alone.

In this thesis, I take an artificial language learning approach to investigate the cognitive underpinnings of word order universals. In Part I, I report three artificial language learning experiments testing a hypothesized link between cognition and cross-category harmony. Cross-category harmony is one of the most well-known statistical typological universals, describing the tendency for word order to correlate across different types of phrases within a language. Explanations for this universal vary as to whether cognitive factors play a role, or instead the tendency is due to mechanisms of language change alone. I investigate two questions: (1) do learners align order across phrase types in a way that corresponds to observed typological tendencies for harmony? (2) what factors exactly determine why only certain phrases exhibit word order harmony? Specifically, I test a learning bias for harmony between two pairs of phrase types: harmony between the verb phrase and adpositional phrase, which is one of the most robust cases of cross-category harmony documented in typology, and harmony between verb phrase and noun phrase consisting of an adjective and a noun, which has been rejected by previous typological work. Across a set of three experiments, I find that learners are biased
in favor of harmony between the order of verb and object and the order of adposition and noun. However, the bias for harmony between the order of verb and object and the order of adjective and noun is weak, and indeed only arises when the semantic similarity between adjectives and verbs is relatively high. When adjectives are active and therefore more verb-like, we find harmony; when they are stative and therefore less verb-like, we do not. These results suggest that the bias for cross-category harmony is not purely based on syntactic notions of head and dependent, often discussed in the literature, but also reflects the interaction between a general cognitive bias favoring consistent order and cross-category similarity.

In Part II, I report three experiments testing whether there is a learning bias for a particular order of numeral, numeral classifier, and noun. Typological research suggests that among the six logically possible ways of ordering these three elements, only the orders in which numeral classifiers are adjacent to numerals are attested. This has been considered one of the most important pieces of evidence distinguishing between alternative analyses of the syntactic status of numeral classifiers: whether the hierarchical structure underlying these phrases is such that numeral classifiers combine with numerals first, or combine with nouns first. However, typological evidence is not very reliable in this case since numeral classifiers are highly geographically constrained and likely reflect language contact. I conduct three artificial language learning experiments to investigate two questions: (1) do learners show a preference for some word order patterns that are predicted by one of the two competing analyses? and (2) does learners’ preference match the tendency we observed in typology? Results show that learners do not exhibit a reliable preference for the word order patterns that are predicted by either of these two hypotheses. I discuss the possibility that the absence of a bias across participants might suggest no universal underlying structure for the phrase consisting of the three elements across languages.

Overall, the evidence presented in this thesis provide mixed evidence for the role of cognition in explaining statistical typological universals. In some cases—like harmony—the distribution of language structures is (at least partly) shaped by the features of human cognitive system. In other cases—like classifier order—however, the distribution of language structures likely reflects other factors, such as
language contact. An artificial language approach helps us to adjudicate between these possibilities, illuminating the role of cognition in shaping particular features of language.
Lay Summary

Human languages exhibit some common patterns in how words are ordered. For example, in some languages, such as English, verbs precede their objects (e.g., *kick the ball*) and adpositions also precede their objects (e.g., *behind the ball*); in other languages, such as Japanese, the order between the two pairs of elements is the opposite: both verbs and adpositions follow their objects. The English and Japanese pattern is very common; it is much less common for languages to have verbs precede their objects and adpositions follow, or the opposite. Word order correlations like these—where two different types of words tend to consistently appear in the same position, first or last, across different languages—have been observed again and again. Nevertheless, order in certain types of phrases appears not to be correlated. For example, in noun phrases such as *blue ball*, the adjective precedes the noun in both English and Japanese. This means that the order of adjectives and nouns does not correlate with the order of verbs and objects, or adpositions and objects. While previous research suggests that consistency in word order across phrases is generally easier to learn, it is not yet clear why some phrases show word order consistency and others don’t.

Another common pattern of word order is found with numeral classifiers. Numeral classifiers are words that are required in numeral constructions in some languages (e.g., in Mandarin: *san ben shu*, lit., three classifier book, ‘three books’). It has been observed that, across languages with classifiers, the numeral and the classifier are always adjacent. However, numeral classifiers are only found in certain regions in the world. Thus it remains a question whether the order of classifiers is driven something special about their meaning or grammatical features, or by the
fact that classifier languages are closely related to each other.

In this thesis, I take an artificial language learning approach to test the cognitive basis of these two word order tendencies. I taught adult participants miniature made-up languages in the lab, to see what they learn, and what word order patterns they produce. I obtained different results in the two cases. In the case of word order correlations, I find that learners have a preference to align verbs in verb phrases with adpositions in adpositional phrases, and a preference to align verbs in verb phrase with adjectives in noun phrases. But the latter only holds when the adjectives express active meanings (e.g., 'broken')—i.e., when they are more similar to the meanings of verbs. When the adjectives are stative such as 'spotty' and 'red', learners have no preference to align them with verbs. These results roughly matches the pattern observed across languages, indicating that word order correlations are (at least partly) shaped by the features of human cognition. Moreover, the results suggest that learners’ bias for consistent order between phrases is sensitive to the similarity of the elements to be aligned. In the second case of numeral classifier order, in all of my experiments, I find no reliable bias for any order between numeral, classifier, and noun, including between the orders that are cross-linguistically frequent and rare. The huge disparity in learning and typology highlights the possibility that the cross-linguistic tendency in the numeral classifier order is likely driven by factors beyond cognition.

Overall, the evidence presented in this thesis shows that cognition plays different roles in shaping cross-linguistic tendencies. An artificial language approach helps us to adjudicate between different possibilities, revealing the (lack of) cognitive foundations of specific features of language.
Acknowledgements

This thesis would not have been possible without the guidance and support from my supervisors Jennifer Culbertson and Simon Kirby. I still remember the first lecture I heard from Jenny, which made me decide to come to Edinburgh. I am really fortunate to have the opportunity to work with these brilliant people in the past years. Jenny is incredibly sharp, insightful, and always there when I faced challenges. Her questions at the supervision meetings have constantly prompted me to think more deeply and thoroughly. I will always remember the edits and comments she made to my drafts, which have been teaching me how to tell a story clearly and concisely. Simon has always reminded me of the big pictures when I was unsure about the next step of my research, as well as spotted the tiny flaws when I conceived the experiments. It’s amazing to me how someone can be both a rigorous scientist and a talented artist at the same time. The guidance and support I received from them made the whole journey a great joy. I am also grateful to Kenny Smith, Marieke Schouwstra, Alex Martin, and Mora Maldonado, who had provided invaluable advice for my work at the supervision meetings during different stages of my PhD studies. I would also like to thank Prof. Danqing Liu from whom I first learnt about word order typology. I really appreciate his continuous encouragement and mentorship over the past many years.

I would also like to thank my peers at CLE: Fausto Carcassi, Tamar Johnson, Andres Karjus, Fiona Kirton, Jonas Nölle, Cathleen O’Grady, Asha Sato, Svenja Wagner, Marieke Wœnsdregt, Henry Conklin, Lauren Fletcher, Shangmin Guo, Maisy Hallam, Annie Holtz, Patrick Kanampiu, Anna Kapron-King, Aislinn Keogh, Yajun Liu, Yevgen Matusevych, Marc Meisezahl, Juan Guerrero Montero, Vlad Nedelcu,
Elizabeth Pankratz, Ponrawee Prasertsom, Sydelle De Souza and Shira Tal. It has truly been a pleasure meeting and working with these wonderful people. I am also thankful to Stefano Coretta and Elizabeth Pankratz for their repeated help and discussions about statistics.

Finally, I want to express my gratitude to my family. To my parents, brother, and sister-in-law, thank you for the love and support you have always provided me and for consistently standing behind me. Thank you to my parents-in-law for their unwavering support and backing for our little family. I am also grateful for my husband for being by my side all along, encouraging me and supporting me in doing what I like. Last but by no means least, special thanks to my little son, for the happiness you’ve brought me. I’m so proud of you.
1 Introduction
   1.1 Typological universals ........................................... 1
   1.2 Explaining typological universals ............................... 2
   1.2.1 Cross-linguistic similarities as results of genealogical relatedness or language contact ........................... 2
   1.2.2 Typological universals as reflections of innate Universal Grammar ........................................... 4
   1.2.3 Typological universals as results of language learning and/or use ........................................... 6
   1.2.4 Typological universals as results of diachronic processes ........................................... 9
   1.3 Linking typological universals and language learning .......... 11
   1.3.1 Typological universals in natural language acquisition ...... 11
   1.3.2 Natural language acquisition: methodological limitations .... 16
   1.4 Artificial language learning ..................................... 17
   1.4.1 Why we need it? ........................................... 17
   1.4.2 Parallels between cognitive biases in artificial language learning and typological tendencies ................. 19
   1.5 Outline of this thesis ............................................ 33

Part I Evidence for the cognitive foundations for cross-category harmony

2 The learning bias for cross-category harmony is sensitive to semantic sim-
3.1.2 Using artificial language learning approach to detect underlying structures ........................................ 96
3.2 Experiment 1 ..................................................................... 97
  3.2.1 Experiment 1A ............................................................. 97
  3.2.2 Experiment 1B ........................................................... 108
3.3 Experiment 2 ..................................................................... 111
  3.3.1 Methods ................................................................... 111
  3.3.2 Results .................................................................... 113
  3.3.3 Discussion ................................................................. 115
3.4 Experiment 3 ..................................................................... 117
  3.4.1 Methods ................................................................... 118
  3.4.2 Results .................................................................... 122
  3.4.3 Discussion ................................................................. 124
3.5 General discussion ............................................................ 127
  3.5.1 Numeral classifiers: cross-linguistically heterogeneous . . . 127
  3.5.2 Disparity in typology and language learning ..................... 131

4 Summary and conclusion ....................................................... 133
  4.1 Summary and contributions ................................................. 133
    4.1.1 Summary of experiments ............................................ 134
    4.1.2 General conclusions ................................................... 138
  4.2 Future directions ............................................................... 139
    4.2.1 Measuring between-category similarity ......................... 139
    4.2.2 Extending to populations with fully non-harmonic native languages ........................................... 140
    4.2.3 Harmony based on alignment of lexical/phrasal categories . . . 141
    4.2.4 The learning bias and processing bias: interaction and competition ........................................... 141
    4.2.5 Incorporating syntactic tests into artificial language learning experiments .................................... 143
    4.2.6 Order of numeral classifiers and gender: the influence of semantic contribution .......................... 144
A  Appendix for Part I 146
   A.1  Learning accuracy in non-critical testing phases 146
      A.1.1  Learning accuracy in non-critical testing phases in Experiment 1 146
      A.1.2  Learning accuracy in non-critical testing phases in Experiment 2 148
      A.1.3  Learning accuracy in non-critical testing phases in Experiment 3 150
   A.2  Second language experience in Experiments 2 and 3 153
   A.3  Learners’ responses in translating artificial words/phrases into their native languages 156
   A.4  The Bayesian model with weakly informative priors in Experiment 3 162

B  Appendix for Part II 164
   B.1  Vocabulary accuracy in non-critical testing phases 164
      B.1.1  Vocabulary accuracy in non-critical testing phases in Experiment 1A and 1B 164
      B.1.2  Vocabulary accuracy in non-critical testing phases in Experiment 2 164
      B.1.3  Vocabulary accuracy in non-critical testing phases in Experiment 3 166
   B.2  Learners’ responses in translating artificial words into their native languages in all experiments 168

C  Appendix for instructions for all experiments 171
   C.1  Instructions for experiments in Part I 171
      C.1.1  Instructions for Experiment 1 171
      C.1.2  Instructions for Experiment 2 174
      C.1.3  Instructions for Experiment 3 175
   C.2  Instructions for experiments in Part II 176
      C.2.1  Instructions for Experiments 1A and 1B 176
      C.2.2  Instructions for Experiment 2 178
List of Figures

2.1 Examples of the visual stimuli in Experiment 1. ........................ 55
2.2 Proportion of responses using harmonic order in Experiment 1. .... 59
2.3 Sample stimuli images for adjectives in Experiments 2 and 3 ......... 65
2.4 Proportion of responses using harmonic order in Experiment 2. .... 67
2.5 Proportion of responses using harmonic order in Experiment 3. .... 76

3.1 Geographic distribution of numeral classifier languages from WACL 95
3.2 Examples of the visual stimuli in Experiments 1A, 1B, and 2. ....... 100
3.3 Average proportion choice of orders predicted by the classifier-with-noun hypothesis in Experiment 1A ................................. 104
3.4 Average proportion choice of N-Num-Cl order in Experiment 1B .... 110
3.5 Average proportion of phrases produced with N-Num-Cl order in Experiment 2 .......................................................... 114
3.6 Examples of the visual stimuli in Experiment 3 ........................... 120
3.7 Average proportion of phrases produced with N-Num-Cl order in Experiment 3 .......................................................... 123

A.1 Proportion of responses using harmonic order across verb phrase training conditions and order types of learners’ second language in Experiment 2. ................................................... 154
A.2 Proportion of responses using harmonic order across verb phrase training conditions and order types of learners’ second language in the adpositional phrase extrapolation condition in Experiment 3 .... 155
B.1 Vocabulary accuracy of nouns, numerals, and classifiers in Experiment 1A ................................................................. 165
B.2 Vocabulary accuracy of nouns, numerals, and classifiers in Experiment 1B ................................................................. 165
B.3 Vocabulary accuracy of nouns, numerals, and classifiers in Experiment 2 ................................................................. 166
B.4 Vocabulary accuracy of nouns, numerals, and classifiers in Experiment 3 ................................................................. 167
List of Tables

1.1 Parallels between typological tendencies and cognitive biases in artificial language learning and communication. .................. 31

2.1 Predicted order of adposition and noun in Experiment 1 ............ 52
2.2 Lexicon of the artificial language in Experiments 1, 2 and 3. ........ 53

3.1 Numeral classifier system in Minangkabau .......................... 90
3.2 Relative typological frequency of different word orders of Num, Cl and N reported in the literature ................................. 92
3.3 Orders predicted based on the two competing hypotheses of interest in Experiment 1A .................................................. 99
3.4 Lexicon of the miniature artificial language in Experiments 1A and 1B. 100
3.5 Counts of different strategies reported by participants in Experiment 1A ................................................................. 106
3.6 Orders predicted based on the two competing hypotheses of interest in Experiment 1B .................................................. 108
3.7 Lexicon of the miniature artificial language in Experiment 2 ........ 111
3.8 Word order variation of the [Num+Cl+N] phrase in some languages. 117
3.9 Lexicon of the miniature artificial language in Experiment 3. .... 119
3.10 Summary of the typological frequency of orders tested and results in each experiment. ................................................. 131

4.1 A summary of all the experiments reported in this thesis. ............ 138

A.1 Model results of noun accuracy in Experiment 1 ................. 147
A.2 Model results of verb accuracy in Experiment 1
A.3 Model results of adposition accuracy in Experiment 1
A.4 Model results of verb phrase vocabulary accuracy in Experiment 1
A.5 Model results of verb phrase order accuracy in Experiment 1
A.6 Model results of verb phrase matching accuracy in Experiment 1
A.7 Model results of noun accuracy in Experiment 2
A.8 Model results of verb accuracy in Experiment 2
A.9 Model results of adposition/adjective accuracy in Experiment 2
A.10 Model results of verb phrase vocabulary accuracy in Experiment 2
A.11 Model results of verb phrase order accuracy in Experiment 2
A.12 Model results of verb phrase matching accuracy in Experiment 2
A.13 Model results of noun accuracy in Experiment 3
A.14 Model results of verb accuracy in Experiment 3
A.15 Model results of adposition/adjective accuracy in Experiment 3
A.16 Model results of verb phrase vocabulary accuracy in Experiment 3
A.17 Model results of verb phrase order accuracy in Experiment 3
A.18 Model results of verb phrase matching accuracy in Experiment 3
A.19 Counts of different types of translations of artificial adpositions by
   English speakers in Experiment 1.
A.20 Counts of different types of translations of artificial adpositions by
   Mandarin speakers in Experiment 1.
A.21 Counts of different types of translations of artificial adjectives by En-
   glish speakers in Experiment 2.
A.22 Counts of different types of translations of noun+color adjective phrases
   in Experiment 3.
A.23 Counts of different types of translations of noun+texture adjective
   phrases in Experiment 3.
A.24 Counts of different types of translations of noun+active adjective
   phrases in Experiment 3.
A.25 Counts of different types of translations of adpositional phrases in
   Experiment 3.
A.26 Parameter estimates in the additional model with informative priors in Experiment 3. ......................................................... 163

B.1 Counts of different types of translations of artificial words in Experiment 1A. ........................................................... 168

B.2 Counts of different types of translations of artificial words in Experiment 1B. ........................................................... 169

B.3 Counts of different types of translations of artificial words in Experiment 2. ........................................................... 169

B.4 Counts of different types of translations of artificial words in Experiment 3. ........................................................... 170
List of abbreviations

1: 1st person
3: 3rd person
ABS: absolutive
AUG: auxiliary
CL: classifier
ERG: ergative
IMPERF: imperfective
REL: relativizer
SG: singular
SFP: sentence final particle
Chapter 1

Introduction

1.1 Typological universals

Human languages show remarkable diversity at almost all levels of the language system (Evans & Levinson 2009). However, it has also been observed that the diversity is not without limitations—some linguistic patterns occur again and again in different languages while some are cross-linguistically rare or even unattested. Since the pioneering work on word order by Greenberg (1963), a great number of typological universals (or tendencies) are generalized based on large scale cross-linguistic comparisons (e.g., Dryer & Haspelmath 2013; Skirgård et al. 2023). Below are two examples of typological universals observed in word order and morphology from Greenberg (1963):

(1) a. Languages with dominant VSO order are always prepositional. With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional. (Universal 3 and 4 in Greenberg 1963)

b. There is no language in which the plural does not have some nonzero allomorphs, whereas there are languages in which the singular is expressed only by zero. (Part of the Universal 35 in Greenberg 1963)

These typological universals posit constraints on the degree to which languages
can vary.\(^1\) For example, \((1a)\) states the constraint on the placement of adpositions in languages with different basic word orders: (verb-initial) VSO languages strongly tend to have prepositions (e.g., Irish) whereas (verb-final) SOV languages strongly tend to have postpositions (e.g., Japanese); and \((1b)\) describes the asymmetric coding of singular and plural meanings across languages: while it is common for languages to have zero marking for singular (e.g., English), zero-marking for plural is far less frequently observed across languages.

Understanding why human languages have these universals is one of the core objectives of linguistics. In the past decades, various types of explanations have been proposed to account for the origin of language universals,\(^2\) including accounts in terms of language relatedness, innate properties particular to the language faculty of the human mind, general cognitive biases, functional pressures, and diachronic processes. In the next section, I will briefly summarize different approaches to explaining typological universals.

### 1.2 Explaining typological universals

#### 1.2.1 Cross-linguistic similarities as results of genealogical relatedness or language contact

Before going to other explanations, a first possibility is that some recurring patterns in different languages may not be the result of universal principles but could instead be products of linguistic relatedness (e.g., Maslova 2000; Ladd et al. 2015;)

---

\(^1\) According to different criteria, typological universals can be divided into various subcategories. For example, based on the modality of universals, typological universals can be categorized as absolute universals (which apply to all known languages) or statistical universals (which apply to most languages); and based on whether the universal is about one linguistic feature or about the relation between different linguistic features, universals can be classified as non-restrictive universals (or unconditional universals, e.g., conditional clauses tend to precede conclusion clauses across languages (Greenberg 1963)) or restrictive universals (or implicational universals, which capture the correlation between two different linguistic features, as shown by \((1a)\)).

\(^2\) Throughout this thesis, "typological universal" and "language universal" are used interchangeably, both of which refer to cross-linguistic tendencies.
Genealogical relatedness and areal relatedness are the two most important sources of data non-independence in typological samples (Ladd et al. 2015). Compared to languages without any genetic relation, languages with a shared historical origin tend to exhibit greater similarity due to their inheritance of shared features from the ancestral language. Likewise, languages in contact are often more likely to exhibit similarities beyond random chance, as they tend to mutually influence each other through processes like borrowing or copying (Thomason 2001; Matras 2010). For example, after long-term contact, genetically unrelated languages in proximity may share a set of language features and form a Linguistic Area.

The issue of data non-independence in language samples has long been noticed by typologists (e.g., Dryer 1989, 1992; Rijkhoff & Bakker 1998; Bickel 2008; Ladd et al. 2015). To mitigate the influence of data non-independence in typological samples, a range of sampling and statistical methods have been put forward, such as stratification sampling (e.g., Dryer 1992; Hammarström & Donohue 2014), diversity sampling (e.g., Rijkhoff & Bakker 1998; Miestamo et al. 2016), hierarchical regression models with genealogical and areal information as predictors (e.g., Cysouw 2010; Guzmán Naranjo & Becker 2022), and phylogenetic methods (e.g., Dunn et al. 2011; Jäger & Wahle 2021) (see also Ladd et al. 2015 and Guzmán Naranjo & Becker 2022). By employing these methods, some cross-linguistic similarities that had been assumed to be universals were later found to be caused by genealogical and areal factors. One example is the relation between verb+object order and noun+adjective order. It had been widely assumed that the order between verb and object is correlated with the order between noun and adjective: VO languages tend to have N-Adj order and OV languages tend to have Adj-N order (Greenberg 1963; Lehmann 1973). However, by using the stratification sampling method, Dryer (1988a, 1992) show that in fact there is likely no such correlation at the global level. Dryer points out that this misunderstanding came from the fact that the two consistent word order patterns are dominant in different parts of the

---

3For example, Bickel (2011: 79) points out that "While the number of proposed statistical universals is impressive — the Universals Archive at Konstanz has collected over 2000 (Plank & Filimonova 2000) —, very few of them have been rigorously tested for independence of area, family, and time."
world: VO languages in southwestern Europe, southeast Asia, sub-Saharan Africa, and Australia tend to have the N-Adj order, and most OV languages in Asia have the Adj-N order. Outside these areas of the world, however, no such patterns are observed. This was also subsequently confirmed by Jäger & Wahle (2021) which uses phylogenetic inference to evaluate word order correlations between different phrases across languages.

Another example is the relation between the verb+object order and affix position. It has been discussed by several researchers that languages with object-verb (OV) order tend to have suffixes while languages with verb-object (VO) order tend to have both prefixes and suffixes (e.g., Greenberg 1957; Bybee et al. 1990; Siewierska & Bakker 1996) (for an overview, see Song 2012: 54-65), but the relation between the two linguistic properties had not been rigorously tested. Guzmán Naranjo & Becker (2022) use hierarchical phylogenetic regression models to assess the potential influence of the verb-object order on affix position. Based on the data reported in Dryer (2013f), Guzmán Naranjo & Becker (2022) compared the results of a model with only verb+object order as a predictor to a model that includes verb+object order, genealogical relations, and areal closeness as predictors. The two models yield different conclusions: the model with verb+object order as the only predictor suggests that the verb+object order is likely to have an effect on affix position, whereas the model with all the three predictors indicates no such effect. Their results demonstrate that the effect of verb+object order on affix position can be overestimated when the sample fails to control for genealogical and areal biases, giving rise to a false correlation between the two linguistic properties.

1.2.2 Typological universals as reflections of innate Universal Grammar

For typological universals which reflect commonalities that are not due to genetic or areal relatedness, there are, of course, a number of different explanations. One explanation is that typological universals are reflections of an innate Universal Gram-
mar (UG) of human language. Under this hypothesis, children are equipped with a universal language faculty (Chomsky 1981) and thus "all languages must be close to identical, largely fixed by the initial state" (Chomsky 2000: 122). For example, in the framework of Principles and Parameters, a child’s a priori language knowledge consists of a number of universal principles that work in all languages and some parameters to be determined by specific languages (Chomsky 1981). One example of this type of explanation is the word order correlation in (1a): the correlation between the adpositional phrase and verb phrase is one of the observable results of the Head-directionality parameter (Travis 1984; Baker 2001), posited within the Principles-and-Parameters framework. The Head-directionality parameter requires a consistent ordering of the syntactic head of a phrase relative to its complement within a language. As a binary parameter, the Head-directionality parameter can only be set to head-initial or head-final in a given language. In head-initial languages, syntactic heads consistently precede their complements across different phrases; whereas in head-final languages, syntactic heads consistently follow their complements. VSO languages are head-initial, with the syntactic head (verb) preceding its complement (object). It is thus natural for adpositions which is the head of adpositional phrase to precede its complement. By contrast, SOV languages are head-final, with the syntactic head (verb) following its complement (object). In adpositional phrases, the head adposition should also follow its complement. The word order correlation between the verb phrase and the adpositional phrase is thus a reflection of the innate Universal Grammar. This type of explanation puts hard restrictions on human language: grammars in all languages are derived from Universal Grammar and thus those that are in conflict with Universal Grammar are biologically impossible. Moreover, this type of explanation typically assumes that the language faculty is one module of human mind and independent of other general cognitive and perceptual abilities (Chomsky 2000). The constraints proposed

4Chomsky (1998: 33): "There has also been very productive study of generalizations that are more directly observable: generalizations about the word orders we actually see, for example. The work of Joseph Greenberg has been particularly instructive and influential in this regard. These universals are probably descriptive generalizations that should be derived from principles of UG."

5However, note that some research within the framework takes a different view. For example, Newmeyer (2004: 527) holds that "most typological generalizations are in no sense ‘knowledge of language’".
in this framework are thus domain-specific (i.e., apply only to human language and not to other cognitive domains).

1.2.3 Typological universals as results of language learning and/or use

An alternative type of explanation contends that typological universals are shaped by general biases of human cognition, including those in language learning, language processing or communication. According to this type of explanation, language structures that are easier to learn, to process, or to produce are more likely to be preferred by individual learners or speakers (e.g., Hawkins 1994, 2004; Christiansen & Chater 2008; MacDonald 2013; Culbertson & Kirby 2016). Through cultural evolution, biases at the individual level could ultimately evolve into typological universals at the population level (e.g., Kirby et al. 2008; Smith 2011; Smith et al. 2017; Blythe & Croft 2021).

One approach within this camp focuses on the role of language learning in the formation of typological universals. A growing body of research proposes that typological universals are (partly) driven by learning biases of individual learners. It has been found that, for a number of language features, the biases exhibited by learners during language learning align with cross-linguistic tendencies (e.g., phonological universals: Martin & Peperkamp 2020; morpho-syntactic universals: Culbertson et al. 2012; Saratsli et al. 2020; Saldana et al. 2021; Maldonado & Culbertson 2022, see Culbertson 2023 and Section 1.4.2 for reviews). For example, it has been argued that the correlation between verb phrase and adpositional phrase word order is driven by a general simplicity bias in the human cognitive system (Culbertson & Kirby 2016). The simplicity bias states that all things being equal, simpler explanations (or hypotheses) are preferred by learners (Chater & Vitányi 2003). Compared to a grammar with different word order rules for different phrase types, a grammar with a consistent order across phrase types is easier and thus preferred by learners. In contrast to the hard domain-specific constraints of innate Universal Grammar, the learning bias explanation makes predictions about the likelihood of a particular grammar. Moreover, this type of explanation underscores the pos-
sibility that biases underlying typological universals are likely to be attributed to the interaction between the domain-general biases and linguistic representations (e.g., Culbertson & Kirby 2016; Ferdinand et al. 2019; Culbertson & Kirby 2022). For example, Culbertson & Kirby (2022) found that in a non-linguistic task (sequences of shapes), learners also preferred consistently placing similar objects in the same position, demonstrating a simplicity bias in the linearization of objects in non-linguistic domain that is parallel to correlations between head and complement order in linguistic phrases.

A second approach centers on language processing or language production, contending that some linguistic patterns are more common because they are easier to process or produce (e.g., Hawkins 1994; MacDonald 2013). For example, Hawkins (1994) provides a processing explanation for word order correlations between different phrases. He argues that structures with a shorter distance between heads of different phrases are easier to process, as the immediate constituents of the head can be identified earlier. Compared to combinations of phrases with both head-initial and head-final orders, combinations that have a consistent head-initial or head-final order generally have shorter dependency distance. One example Hawkins provides is the combination of verb and adpositional phrases. There are four logically possible structures: (i) \( VP[PP[Prep N]] \), (ii) \( VP[PP[N Post] V] \), (iii) \( VP[Prep N] V \), (iv) \( VP[VPP[N Post]] \). In the first two structures, both the order between the verb and adpositional phrase and the order between the adposition and noun is head-initial (for i) or head-final (for ii), whereas in the last two

6 Hawkins (1994: 77) formulates this as “Early Immediate Constituent (EIC)”: “I believe that words and constituents occur in the orders they do so that syntactic groupings and their immediate constituents (ICs) can be recognized (and produced) as rapidly and efficiently as possible in language performance. Different orderings of elements result in more or less rapid IC recognition”. The IC-to-word ratio is calculated by dividing the number of ICs in the domain by the total number of words in that domain. For example, the IC-to-word rations of the following four sentences are:

(1) Joe looked up the number. (IC-to-word ratio: 3/3 = 100%)
(2) Joe looked the number up. (IC-to-word ratio: 3/4 = 75%)
(3) Joe looked the number of the ticket up. (IC-to-word ratio: 3/7 = 42.9%)
(4) Joe looked the number that Mary had forgotten up. (IC-to-word ratio: 3/8 = 37.5%)

The verb phrase in these sentences has three ICs: the verb \( \text{looked} \), the particle \( \text{up} \), and the noun phrase. The IC-to-words ratios of the four sentences are shown above.
structures, the orders of the two types of phrase are inconsistent (head-final order in the verb phrase but head-initial order in the adpositional phrase in iii, and head-initial order in the verb phrase but head-final order in the adpositional phrase in iv). From the perspective of processing, the first two structures are optimal because the syntactic head verb and adposition are adjacent and the domain of VP can be recognized earliest. By contrast, the last two structures are not the optimal: the noun is inserted between the adposition and verb which results in the delay of the recognition of the VP domain. MacDonald (2013) brings in both simplicity and processing explanations and argues that the distributional regularities of language features are driven by a pressure to minimizing production difficulty. For example, she argues that common word order patterns are generally motivated by three principles to reduce cognitive burdens in working memory and utterance planning: putting easier things earlier (e.g., subject before object), reusing word order patterns (e.g., word order rigidity) and reducing interference (e.g., a tendency for using passive structures (to demote agent) to reduce interference when the agent and patient are semantically similar).

A third approach focuses on language use, arguing that cross-linguistic tendencies are the results of an interaction between competing functional pressures (e.g., Haiman 1983; Comrie 1989; Bybee 2010; Haspelmath 2021). On the one hand, language needs to be as clear as possible to facilitate successful communication. On the other hand, language needs to be as economical as possible to facilitate learning and/or use. Language universals emerge as a result of balancing the two forces. This has long been discussed by linguists, though the concepts and terms involved in the literature vary, such as laziness (Bequemlichkeit) and clarity (Deutlichkeit) (Gabelentz 1901) (translations cited from Haspelmath 2014), unification and diversification (Zipf 1949), iconicity and economy (Haiman 1983; Croft 2002), efficiency and complexity (Hawkins 2004), predictability and coding efficiency (Haspelmath 2021). For example, it has been argued that the asymmetric coding of singular and plural meanings is motivated by communicative efficiency: languages tend to use non-zero marking for meanings that are less frequent and thus more likely to be ambiguous. Compared to singular meanings, plural meanings are less frequent and in greater need of disambiguation (e.g., Greenberg 1963; Croft 2002; Haspelmath &
The view that typology is shaped by communicative efficiency has also recently been supported by computational and experimental evidence (e.g., Gibson et al. 2019; Hahn et al. 2020; Mollica et al. 2021; Hahn & Xu 2022; Kurumada & Grimm 2019). For example, Kurumada & Grimm (2019) showed that when English speakers learnt a new language with optional plural marking, they tended to condition the use of plural marking based on the frequency of plural meanings of different nouns: learners were less likely to produce plural marking for nouns that are more likely to occur with plural meaning, but more likely to produce plural marking for nouns that are less likely to occur with plural meaning.

1.2.4 Typological universals as results of diachronic processes

In contrast to the cognition-internal explanations reviewed above, there are also explanations in terms of diachrony, arguing that typological universals at the synchronic level are not shaped by preferences in language learning or use, but are products of historical changes at some point during language evolution (e.g., Givón

---

7 Note that while the functional explanation suggests that the coding asymmetry results from the interaction of competing functional pressures, the specific pressures involved vary between different viewpoints. For example, one view holds that this coding asymmetry is caused by the degree of markedness of different values of grammatical categories (e.g., Greenberg 1963; Croft 2002). Compared to singular, plural is semantically more marked (or more complex) and thus more likely to be structurally coded (for the other characteristics of typological markedness such as the marked value tends to have less morphological distinction (e.g., gender distinction in singular pronouns is more common than gender distinction in plural pronoun across languages, such as English) and lower frequency in text, see Croft 2002: 87-101). An alternative view holds that this coding asymmetry is caused by differences in frequency or predictability of meanings (e.g., Haspelmath 2021, 2023). For nouns that are less likely to occur in the plural (e.g., man), the plural meaning of the nouns are less predicted (more surprising) and thus more likely to be marked; however, for nouns that are more likely to occur in the plural (e.g., pea), the singular meaning of the nouns are less predicted (more surprising) and thus more likely to be marked.

8 Note that in the experiments in Kurumada & Grimm (2019), visualization of different nouns are different: collective nouns that dominantly appeared with plural meanings are small insects (appearing individually or in group), whereas individual nouns that dominantly appeared with singular meanings are big animals (appearing individually or in two). As pointed out by the authors, the results could reflect learners' semantic knowledge of the visual referents in their native language. To determine whether this was the case, they conducted a follow-up experiment in which the associations between the noun classes (animals vs. insects) and visual features of the referents (i.e., size) are reversed. No difference in using plural markings between the two noun classes was observed, suggesting that the use patterns of plural markers is also affected by the visual features of the nouns.
1975; Aristar 1991; Bybee 2010; Collins 2019; Cristofaro 2019). For example, there is an alternative diachronic explanation for the alignment of verbs and adpositions in word order: verbs are a common grammatical source of adpositions (e.g., Givón 1975; Aristar 1991). As the linear order between verbs and nouns is likely to remain stable during the grammaticalization process, adpositions naturally maintain the position of verbs, leading to the word order correlation between the verb phrase and adpositional phrase (e.g., Givón 1975; Bybee 1988; Aristar 1991). Regarding the asymmetric coding of singular and plural across languages, there are also explanations from a diachronic perspective. For example, Cristofaro (2019) argues that diachronic data across languages demonstrate that this universal is not due to functional pressures, but a result of various diachronic processes (from different source structures and through different mechanisms). For instance, in some languages both singular and plural were initially zero marked, but later another expression that co-occurred with the plural (e.g., distributive expressions or expressions of multitude in some languages) evolved into a plural marker, leading to the asymmetric coding of singular and plural marking; there are also languages in which both singular and plural were originally zero marked, but later some other expressions that co-occurred with singular evolved into singular marker, leading to the result of non-zero marking of singular but zero marking of plural; in other languages, both singular and plural were originally non-zero marked, but sound changes lead to the loss of singular markers. Given that in some cases, the diachronic changes can be observed directly, diachronic explanations are sometimes considered the "cheapest", or most parsimonious explanations for typological universals (Haspelmath 2019a). For example, when discussing word order correlation between verb and object and auxiliary verb and verb, Bybee (2010: 111) points out that "[f]or this case, no synchronic principles (such as Cross-Category Harmony (Hawkins 1983)) are necessary; grammaticalization gives us the correct orders for free."

Undoubtedly, the factors mentioned in all of the different explanations summarized above are likely to contribute to some degree in shaping the distribution of language patterns. For example, while language contact itself is a cognition-
external factor, the diffusion of language patterns during language contact, may still be driven by some universal cognitive principles: all things being equal, structures that are preferred in terms of learning or processing might be more likely to spread during language contact compared to less preferred patterns (see discussion in Culbertson et al. 2012 and Bickel 2017). Similarly, for cognitive/functional and historical explanations, while the claims of the two approaches seem to be incompatible, it has been pointed out by a number of researchers that both types of accounts are in fact needed in explaining universals (e.g., Kiparsky 2008; Moravcsik 2013; Dryer 2019; Schmidtke-Bode & Grossman 2019).9 The central question is the explanatory role of different hypotheses (Kiparsky 2008), or in other words, to what extent different forces shape a typological universal (Bickel 2017). In this thesis, I focus on the role of language learning in shaping typological universals and ask whether there are cognitive underpinnings for certain typological universals in word order. Throughout, however, I will discuss the potential contributions of other factors, and how exactly they might relate to learning.

1.3 Linking typological universals and language learning

1.3.1 Typological universals in natural language acquisition

Given the focus of this thesis on the role of learning in explaining language universals, it is worth discussing in more detail what previous work has show regarding this connection. Studies on natural language acquisition have, in some cases, demonstrated patterns in language learning that are, to some extent, consistent

9For example, Kiparsky (2008: 27) points out that “historical explanations, once spelled out, often turn out to appeal implicitly to tendencies that are themselves in need of explanation.” The implication here is that these ultimate explanations may be cognitive in nature. Moravcsik (2013: 251) also argues that there are different levels of explanations, though they may differ on the ultimate motivation of universals:” (a) Crosslinguistically common structural patterns are explainable by crosslinguistically common historical changes. (b) Crosslinguistically common historical changes are explainable by crosslinguistic patterns of language acquisition and language use. (c) Crosslinguistic patterns of language acquisition and language use are explainable by language function: the goals of language use and the conceptual tools that are conducive or necessary for it.”
with typology (e.g., Jakobson 1968; Clark 1976; Luján et al. 1984; Clark 2001; van Lier 2005; Diessel & Tomasello 2005). For example, Jakobson (1968) is one of the earliest works to note the correspondence between patterns observed in children’s first language acquisition and typology. He found that the order of acquisition of phonemes is similar to a set of tendencies observed across languages in the world. For example, the existence of fricatives implies the existence of stops across languages. Similarly, children typically acquire stops at an earlier developmental stage than fricatives. In later work, Hawkins (1987) proposes the Universal Consistency Principle, arguing that any stage in first language acquisition and second language acquisition should be consistent with universals derived from synchronic evidence. Below I will take a well-known typological universal, the Noun Phrase Accessibility Hierarchy, as an example and review some representative studies that explore whether this universal is manifested in first and second language acquisition. 10

The Noun Phrase Accessibility Hierarchy is a typological generalization regarding the accessibility of relativization of nouns in different grammatical roles. Based on a sample of roughly fifty languages, Keenan & Comrie (1977) found that while different languages exhibit variation on what grammatical roles can be relativized and strategies used for relativization of different grammatical roles, they tend to conform to a universal hierarchy, which is shown in (2):

(2) Accessibility Hierarchy (AH) (Keenan & Comrie 1977: 66)
Subject > Direct object > Indirect object > Oblique > Genitive > Object of comparison. (”>” means more accessible than)

The hierarchy is interpreted as follows (Keenan & Comrie 1977: 67): "(i) A language must be able to relativize subjects; (ii) Any RC-forming strategy must apply to a continuous segment of the AH; (iii) Strategies that apply at one point of the AH may in principle cease to apply at any lower point." These constraints imply that if a grammatical role on the hierarchy can be relativized, then all the grammatical roles that are higher on the hierarchy can also be relativized. For example, for a given

10For the comprehensive review of typological universals in first language acquisition, see Bow- erman (2011), and for that in second language acquisition, see Eckman (2011).
language, if the indirect object can be relativized, then the subject and direct object in this language can also be relativized.\textsuperscript{11}

Keenan & Comrie (1977) argue that this typological generalization reflects psychological ease. Data from child language acquisition in some languages suggest that the pattern of children’s acquisition of relative clauses aligns with the AH. For example, Diessel & Tomasello (2001) analyzed the data of four English-speaking children (aged 1;9-5;2) from the CHILDES corpus, and found that the frequencies of different types of relative clauses conform to the AH: frequency of subject relative clauses is higher than that of direct object relative clauses, which is in turn, higher than that of oblique relative clauses. Diessel & Tomasello (2005) tested four-year-old English- and German-speaking children’s knowledge of relative clauses using a sentence repetition task. Four types of relative clauses were tested in their experiments: Subject (including subject of intransitive and transitive verbs) relative clauses, direct object relative clauses, indirect object relative clauses, and genitive relative clauses. They found that despite the differences between relative clauses in English and German, the acquisition patterns in the two languages are quite similar: subject relative clauses have a lower error rate compared to direct object relative clauses, which, in turn, have a lower error rate than indirect and oblique relative clauses; lastly, genitive relative clauses exhibit the highest error rate among all the different types of relative clauses.

Data from some second language acquisition studies also confirm the predictions of AH. The AH not only makes predictions on accessibility of relativization of

\textsuperscript{11}English examples of the relativization of different grammatical roles are shown below (examples are cited from Eckman (2011: 622), with the underlines that indicate the relative clauses added here):

a. Subject relative clause: There is the woman who is my sister.

b. Direct object relative clause: There is the woman who(m) I registered.

c. Indirect object relative clause: There is the woman to whom I sent an application.

d. Oblique relative clause: There is the woman about whom I read in the newspaper.

e. Genitive relative clause: There is the woman whose sister graduated last year.

f. Object of comparison relative clause: There is the woman who I am older than.
different grammatical roles, but also strategies used for the relativization of these roles. One type of relativization strategy uses resumptive pronouns in the relative clause to express which grammatical role is being relativized. Keenan & Comrie (1977) show that resumptive pronouns are more likely to be used for relative clauses at the lower positions than those at the higher positions on the hierarchy. These predictions of AH are in general supported by a number of studies on second language acquisition (e.g., English L2 learners: Gas 1979; Italian L2 learners: Croteau 1995; Chinese L2 learners: Xu 2014, also see reviews in Song 2002). For example, Gas (1979) tested L2 English learners with different native languages (i.e., Arabic, Chinese, French, Italian, Korean, Persian, Portuguese, Japanese and Thai) on two tasks: a acceptability judgement of English relative clauses and a task of combining two simple English sentences into a complex sentence with a relative clause. She found that participants’ error rate decreased from the highest position to the lowest position on the hierarchy except for genitive relative clauses. Furthermore, while learners whose native language employs a pronoun-retention strategy for relative clauses were more likely to use this strategy compared to those whose native language has no such strategy, learners tended to use resumptive pronouns for lower positions regardless of whether their native language had such a strategy. These results suggested that while the L2 learning of relative clauses is to some extent influenced by learners’ native language, the learning still complies with the universal constraints found across languages.

However, results from some research in first and second language acquisition cannot be fully explained by the AH. For example, Rahmany et al. (2011) tested Persian-speaking children’s (2;6-7;5) comprehension of subject relative clauses, object relative clauses, and genitive relative clauses using a picture-selection task. Persian is an SOV language with postnominal relative clauses. They found that object relative clauses and genitive relative clauses were more difficult for children to comprehend than subject relatives, which is predicted by the AH. However, there was no difference in the performance on object relative clauses and genitive relative clauses. Since object and genitive relative clauses both have non-canonical order in Persian, the authors interpreted the results not in terms of the AH but instead as evidence that children have particular difficulty in processing sen-
sentences with non-canonical word order. Gutierrez-Mangado (2011) tested Basquespeaking children’s knowledge of relative clauses. Basque is an SOV ergative-absolutive language with prenominal relative clauses. They found that comprehension accuracy of object relative clauses was higher than subject relative clauses. These results are contrary to the prediction of the AH, but could be explained by the morphological markedness of the different types of relative clauses and the linear-distance hypothesis (Hawkins 1994). In Basque, both object relative clauses and subject relative clauses use a gap strategy. There could be a mismatch between the case in the gap (i.e., ergative) and the case of head noun in subject relative clauses (i.e., absolutive), but there is no such mismatch in object relative clauses (i.e., absolutive). The mismatch of case in subject relative clauses could in principle lead to processing difficulties. Moreover, the linear length between the gap and head noun is longer in subject relative clauses than that in object relative clauses. According to the linear-distance hypothesis, structures with longer distance between the gap and head noun are more difficult to parse. This could also increase the processing difficulty of subject relative clauses compared to object relative clauses.

Research on Persian- and Basque-speaking children demonstrates that in addition to the accessibility of grammatical roles, some other cognitive mechanisms, like a bias for canonical word order and a preference for shorter linear distances between the gap and the head noun, may also come into play in learning and processing relative clauses. Because languages exhibit variation in the relevant syntactic structures, such as basic word order, the order between relative clauses and nouns,

---

Below are two examples of subject relative clauses (a) and object relative clauses (b) showing these differences in Basque (Gutierrez-Mangado 2011: 180):

(i) a. *Hau da [e_i amona muxukatzen duen] neska_i.*
   this is grandmother-ABS kiss-imperf aux-rel girl-ABS
   ‘This is the girl who is kissing the grandmother.’

   this is grandmother-ERG kiss-imperf aux-rel girl-ABS
   ‘This is the girl who the grandmother is kissing.’
and their relativization strategies, the role of these different mechanisms may also vary across languages, giving rise to different learning patterns (see discussions in Song 2002).  

1.3.2 Natural language acquisition: methodological limitations

While natural language acquisition data serve as a critical source of evidence for evaluating universal constraints on human languages, this approach comes with its limitations (Fedzechkina et al. 2016; Culbertson 2023). First, natural language acquisition data are likely to be confounded by various factors. For example, in natural language acquisition, it’s hard to control the frequency of language structures. Since language learning is sensitive to the frequency of structures in the input (e.g., Yang 2004; Tomasello 2005), it is possible that order of acquisition, or differences in production or comprehension accuracy merely reflect the frequencies of different structures in the input. For instance, the observation that English-speaking children’s knowledge conforms to the AH cannot be easily disentangled from the fact that the frequencies of different types of relative clauses follow the AH in children’s input. Corpus studies of English do show that the frequency of subject relative clauses is higher than that of object relative clauses: Gordon & Hendrick (2005) found that across three corpus (i.e., Brown, Switchboard, CHILDES), the frequency of subject relative clauses is higher than that of object relative clauses, which in turn is higher than oblique relative clauses; Reali & Christiansen (2007)’s investigation of a larger corpus also confirms that subject relative clauses occur more frequently than object relative clauses in English. Given that typologically unmarked features and constructions are likely to have higher frequencies (e.g., Croft 2002; Moravcsik 2013), it’s problematic to use natural acquisition data alone to posit causal relationships between language learning and typological universals.

Second, it is challenging to test whether a cognitive bias exists in different populations based only with data from natural language acquisition. For example, as

---

13For example, Ozeki & Shirai (2007: 169) point out that “[a]lthough Keenan and Comrie’s (1977) noun phrase accessibility hierarchy (NPAH) has been shown to predict the difficulty order of relative clauses (RCs) in SLA (second language acquisition), most studies of the NPAH have been on European languages.”
shown in the review above, learning (and processing) of relative clauses might be influenced by different cognitive mechanisms. Due to variations in specific language structures, the roles of different cognitive strategies may also vary across languages, resulting in different acquisition patterns and differences in alignment with the prediction of the AH. For example, in English, the AH and the shorter linear distance hypothesis both predict that subject relative clauses should be easier for learners than object relative clauses, it is thus not surprising to observe the preference for subject relative clauses in language acquisition. However, in Basque, the AH and the shorter linear distance hypothesis contradict with each other and the former is likely to be overridden by the latter, leading to an opposite pattern from English.

Another limitation lies in the impracticability of using natural acquisition data for comparing the learning of attested and unattested structures. For example, it has been argued that a hypothetical rule that the first and last segment in a word should agree in terms of phonological features (i.e., First-Last Assimilation) is currently unattested across languages (Lai 2012). It is debated whether the absence of such a rule is due to learnability constraints, e.g., in terms of computational complexity (Heinz 2010; Lai 2012; Finley 2012). Since these patterns are not found in any languages, it is impossible to adjudicate between different explanations using natural acquisition data.

1.4 Artificial language learning

1.4.1 Why we need it?

To deal with these challenges, artificial language learning experiments have been used to investigate the relationship between language learning and typological universals. In such experiments, learners are exposed to (part of) an artificial language or language-like system in a controlled laboratory setting, and are later tested on

---

14 For instance, for a language with first-last consonant voicing pattern, [boteg] would be a possible word because the first consonant ([b]) and the last consonant ([g]) are both voiced; however, *[boget]* would not be a possible word because the first consonant ([b]) is voiced and the final consonant ([t]) is voiceless (Finley 2012: 1536).
their knowledge of the language. Since artificial languages are precisely designed for the structure of interest, causal relations between language structures and language learning can be directly tested (e.g., Fedzechkina et al. 2016; Culbertson 2023). For instance, Hudson Kam & Newport (2005) showed that when exposed to an artificial language featuring a type of irregularity that is rare in natural language, learners adapted the language to make it more systematic (more akin to natural language). This approach enables us to directly assess whether typological universals are shaped by the learnability of different patterns. In other words, all things being equal, this approach allows us to ask whether typologically more frequent patterns are easier to learn than those that are less frequent. For example, Tabullo et al. (2012) compare the learnability of four basic word order patterns between subject, verb, and object: SVO, SOV, VSO, and OSV. SVO and SOV orders are typologically frequent orders which represent roughly 85% of languages in the world, whereas VSO and OSV are typologically infrequent orders which just represent less than 10% of human languages (Dryer 2013e). They taught Spanish-speaking adults a miniature language featuring one of the four word order patterns and asked learners to classify new sentences as “correct” or “incorrect” based on what they had learned during training. They found accuracy rates in the SVO and SOV conditions were higher than those in the VSO and OSV conditions, roughly reflecting the pattern in typology. Given that the input that learners received in different conditions was exactly the same except the feature of basic word order, their results provide evidence for a hypothesized link between learnability and typological patterns of basic word order. However, it is worth noting that since Spanish is an SVO language, the higher accuracy in learning SVO may simply be because learners were better at learning a native-like language compared to others. This brings up the question how to mitigate the influence of learners’ native language experience in artificial language learning experiments. I will come back to this just below.

Moreover, artificial language learning experiments allow us to compare the attested versus unattested patterns. For example, due to its absence in human languages, no acquisition data is available to test whether the First-Last Assimilation mentioned above is avoided due to the learnability. Artificial language learning ex-
periments, however, can help to determine whether this is the case. In fact, there is evidence from artificial language learning experiments suggesting that this rule can be learnt by learners to some degree, though learners were less sensitive to this rule compared to other phonological rules (Finley 2012; Avcu & Hestvik 2020). In addition, artificial language learning experiments also allow us to test whether cognitive constraints on human language are hard (i.e., inviolable) or soft (i.e., violable), and whether the cognitive mechanisms that shape language are domain-specific or domain-general, offering direct behaviour evidence to disentangle different explanations for language universals (Culbertson 2023).

It is worth noting that this experimental framework also has potential limitations. For example, in some cases, it can and should be used across populations with different language backgrounds. This is critical for assessing whether biases observed in one language group are universal or a reflection of their experience with their native language. For example, Culbertson et al. (2012), Culbertson & Newport (2015) and Culbertson et al. (2020a) tested different populations and found that learners tended to put different modifiers to the same side of noun (harmonic order) regardless of whether the order of the noun phrase in their native language is harmonic or non-harmonic, suggesting that the bias is likely to be universal. I will discuss this further in Chapter 2, where I take a similar approach.

1.4.2 Parallels between cognitive biases in artificial language learning and typological tendencies

In the last few decades, there has been a remarkable rise in the use of artificial language learning experiment to investigate the relation between the learnability of different structures and their typological distributions. A number of parallels between cognitive biases in artificial language learning and typological tendencies have been observed. Table 1.1 summarizes these parallels in morphology and syntax (see also Culbertson 2023 for a detailed review of artificial language learning in syntax, and Moreton & Pater 2012a,b for a review of the artificial language learning work in phonology).
<table>
<thead>
<tr>
<th>Linguistic topic</th>
<th>Typological generalization/pattern</th>
<th>Cognitive bias in artificial language learning or communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic word order</td>
<td>SOV order is the most frequent basic word order across languages (Dryer 2013e).</td>
<td>Non-signing adult participants with different native languages were shown pictures of transitive events and were asked to convey the meanings of these pictures with their hands (the silent gesture paradigm). Participants tended to produce Actor-Patient-Act (SOV) order (Goldin-Meadow et al. 2008; Futrell et al. 2015a).</td>
</tr>
<tr>
<td></td>
<td>SOV and SVO are the two most frequent basic word orders across languages (Dryer 2013e).</td>
<td>English-speaking adults were taught an artificial language with one of the six basic word orders and were tested on comprehension (choosing the correct video for a sentence in the language) and production (describing transitive event videos with this language). Participants showed higher production accuracy in SOV and SVO languages than in languages with the other four word order patterns (Tily et al. 2011). Spanish-speaking adults were taught an artificial language with either SVO, SOV, OSV, or VSO order and were tested on a grammatical judgement task. Participants showed higher comprehension accuracy and shorter response time in SOV and SVO languages than in VSO and OSV languages (Tabullo et al. 2012).</td>
</tr>
<tr>
<td>Linguistic topic</td>
<td>Typological generalization/pattern</td>
<td>Cognitive bias in artificial language learning or communication</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>Word order in noun phrase (NP)</td>
<td>When the descriptive adjective precedes the noun, the demonstrative and the numeral, with overwhelmingly more than chance frequency, do likewise (Greenberg 1963).</td>
<td>Adult and child learners with different native languages were taught an artificial language with variable word order between nouns and adjectives and between nouns and numerals (only one modifier for each phrase). They were tested on the relative order between adjectives and numerals. Participants tended to regularize the input towards harmonic patterns (i.e., modifiers (adjective, numeral) are at the same side of the head noun). This harmony bias was replicated in different age groups of different populations whose native languages are either harmonic or non-harmonic in terms of the order in NP (Culbertson et al. 2012; Culbertson &amp; Newport 2017; Culbertson et al. 2020a).</td>
</tr>
</tbody>
</table>

When any or all of the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in that order. If they follow, the order is either the same or its exact opposite (Greenberg 1963). Adult learners with different native languages were taught noun phrases with different single modifiers (adjectives, numerals, or demonstratives, all at the same side of the noun), and were tested on the relative order between multiple modifiers. Learners tended to prefer homomorphic orders—those which preserve the hierarchical structure of the noun phrase (e.g., N-Adj-Num-Dem order in languages with postnominal modifiers and Dem-Num-Adj-N order in languages with prenominal modifiers) in artificial language learning (Culbertson & Adger 2014; Martin et al. 2019, 2020). The bias for homomorphic orders was replicated in iconic artificial language learning (Shapiro & Steinert-Threlkeld 2023) and silent gesture experiments (Culbertson et al. 2020b).

---

15 The hypothesized hierarchical structure of the noun phrase is [Dem [Num [Adj N]]]. Homomorphic orders are the orders in which the distance between adjectives and nouns is not greater than that between numerals and nouns and that between demonstratives and nouns, and the distance between numerals and nouns is not greater than that between demonstratives and nouns.
<table>
<thead>
<tr>
<th>Linguistic topic</th>
<th>Typological generalization/pattern</th>
<th>Cognitive bias in artificial language learning or communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word order in noun phrase (NP)</td>
<td>Genitives tend to precede nouns across languages (Croft 2002; Dryer 2013b).</td>
<td>English-speaking adults were shown pictures featuring an owner and an object and were asked to select between two gesture videos the one that they considered the best for describing the image. Participants showed a preference for the video that started by signing the owner first and then the object (Genitive-Noun order) over the one that began by signing the object first and then the owner (Noun-Genitive order) (Holtz et al. 2022).</td>
</tr>
<tr>
<td>(Modifying) adjectives tend to follow nouns across languages (Dryer 2013a).</td>
<td>(Modifying) adjectives tend to follow nouns across languages (Dryer 2013a).</td>
<td>English-speaking adults were shown pictures of an object with specific property, and were asked to select between two gesture videos the one that they considered the best for describing the image. Participants showed a preference for the video that started by signing the object first and then the property (Noun-Adjective order) over the one that began by signing the property first and then the object (Adjective-Noun order) (Holtz et al. 2022).</td>
</tr>
<tr>
<td>Word order between negative morpheme (Neg) to precede verb (V) across languages (Dryer 1988b, 2013c).</td>
<td>A preference for negative morpheme (Neg) to precede verb (V) across languages (Dryer 1988b, 2013c).</td>
<td>English-speaking adults were exposed to an artificial VSO language with either dominant NegV order, dominant VNeg order, or both NegV and VNeg orders with equal frequency. More NegV order than that in the input was produced in all three types of languages, suggesting that participants had a preference for NegV order (Burgess 2022).</td>
</tr>
<tr>
<td>Linguistic topic</td>
<td>Typological generalization/pattern</td>
<td>Cognitive bias in artificial language learning or communication</td>
</tr>
<tr>
<td>--------------------</td>
<td>-----------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>Morpheme ordering</td>
<td>Where morphemes of both number and case are present and both follow or both precede the noun base, the expression of number almost always comes between the noun base and the expression of case (Greenberg 1963).</td>
<td>English- and Japanese-speaking adults were taught an artificial language featuring nouns with number and case morphemes (both at the same side to the noun stem, only one morpheme in each noun), and were tested on the relative order between the two morphemes. Participants tended to generalize orders in which number morphemes are closer to noun stems than case morphemes (Saldana et al. 2021).</td>
</tr>
<tr>
<td>Affix ordering</td>
<td>Languages have a preference for suffixes over prefixes (Greenberg 1957; Hawkins &amp; Gilligan 1988; Hawkins &amp; Cutler 1988)</td>
<td>English-speaking adults were exposed to sequential syllables, shapes, and musical notes with differences either at the ends or at the beginnings. Sequences with differences at the ends were considered more similar than those with differences at the beginnings (Hupp et al. 2009); English-speaking adults were taught two categories of artificial words with either prefixes or suffixes. Participants showed higher accuracy for the identification of the category with suffixes than that with prefixes (St. Clair et al. 2009). However, Martin &amp; Culbertson (2020) replicated Hupp et al. (2009)’s experiments on native Kîîtharaka-speaking adults whose native language is dominantly prefixing and found an opposite pattern: sequences with differences at the beginnings were considered more similar to Kîîtharaka-speaking adults, calling into question whether this cross-linguistic tendency is shaped by universal features of cognitive system.</td>
</tr>
<tr>
<td>Linguistic topic</td>
<td>Typological generalization/pattern</td>
<td>Cognitive bias in artificial language learning or communication</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>Affix position</td>
<td>Inflectional affixes expressing the same grammatical category tend to appear in the same morphological position in the word (Mansfield et al. 2020).</td>
<td>English-speaking adults were taught an artificial language with noun stems for shapes and suffixes for color and number features of the shape. The language had either fixed position for each suffix category (simple grammar), or fixed position for each suffix, but different positions for different suffixes of the same category (complex grammar), or random position for each suffix (random grammar). Learners were tested on a production task (selecting the correct label to describe an image). Learners’ production accuracy was highest on the simple grammar compared to the other two grammars (Mansfield et al. 2022).</td>
</tr>
<tr>
<td>Reduplication patterns</td>
<td>Base-dependent partial reduplication is extremely rare across languages (Haugen &amp; Hicks Kennard 2011).</td>
<td>English-speaking adults were trained on an artificial language with full reduplication expressing plural meaning and one of three types of partial reduplication expressing augmentative meaning: a CV- partial reduplication, or a CVC- partial reduplication, or a base-dependent CVX partial reduplication. They were tested on a forced-choice task for correct plural and augmentative forms. Participants showed lower accuracy with CVX-base-dependent partial reduplication compared to CV- and CVC- partial reduplication and full reduplication (Haugen et al. 2022).</td>
</tr>
</tbody>
</table>

---

16 Base-dependent reduplication is a type of partial reduplication in which the syllable structure of the reduplicant is determined by the syllable structure of its base. For example, in Mokilese, to express the progressive meanings, verbs with consonant-initial stems need to have a partial CVC reduplicant as the prefix (e.g., *pɔdɔk* ‘to plant’ > *pɔdpɔdɔk* ‘to be planting’), whereas verbs with vowel-initial stems need to have a VCC reduplicant as the prefix (e.g., *andip* ‘to spit’ > *andandip* ‘to be spitting’) (Harrison 1976; Haugen et al. 2022: 300).

17 Full reduplication is the form in which the reduplicant is the same as the base (e.g., *vamseta* ∼ *vamsetavamseta*); CV- partial reduplication is the form with a CV- reduplicant (e.g., *vamseta* ∼ *vvamseta*); CVC- partial reduplication is the form with a CVC- reduplicant (e.g., *vamseta* ∼ *vvamtamseta*). CVX- partial reduplication is either CV- or CVC- form, depending on the syllables of the stem (Haugen et al. 2022: 302).
<table>
<thead>
<tr>
<th>Linguistic topic</th>
<th>Typological generalization/pattern</th>
<th>Cognitive bias in artificial language learning or communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependency</td>
<td>Branching dependencies are generally more prevalent than center-embedded dependencies across languages (Hawkins 2004: 128-129); crossed dependencies is rare in natural languages (Hunter 2021)(^\text{18}).</td>
<td>English-speaking adults were taught noun phrases featuring branching structure, embedded structure or crossed structure and tested on both a comprehension and a production task. Participants showed higher accuracy in both tasks for phrases with branching structure compared to those with embedded structure and crossed structure (Davis &amp; Smith 2023).</td>
</tr>
</tbody>
</table>

---

\(^{18}\)Example of branching dependency, center-embedded dependency, and crossed-dependency are shown below (Vosse & Kempen 1991: 74):

- **English:** when John saw Peter walk
  - subj
  - subj

- **German:** ...als Johan Peter laufen sah
  - subj
  - subj

- **Dutch:** ...toen Jan Peter sag lopen
  - subj
  - subj

---

25
Dependency length minimization is a universal quantitative property of human languages (Futrell et al. 2015b). English-speaking adults were taught either a verb-initial language or a verb-final language (both with object case marking) and were tested on how they ordered subject and object (which was a complex noun) in this language. Participants showed a bias for minimized dependency length between verbs and arguments in both languages. Specifically, they showed a preference for SO order in verb-initial language (e.g., V[punch] S[chef] O[referee next to blue skateboard]), and a preference for OS order in verb-final language (e.g., O[blue skateboard next to referee] S[chef] V[punch]) (Fedzechkina et al. 2018).

Nouns that are more likely to occur with plural meanings tend to have singular marking whereas those are less likely to occur with plural meanings tend to have plural marking (Haspelmath & Karjus 2017). English-speaking adults were exposed to a language with collective nouns that dominantly appeared with plural meanings and individual nouns that dominantly appeared with singular meanings. 66% of all plural meanings of both types of nouns in the language were marked. Participants tended to condition the plural marking based on the predictability of plural meanings: they tended to produce plural marking on individual nouns which were less likely to occur with plural meanings compared to collective nouns (Kurumada & Grimm 2019).

---

19For the verb-initial language, the overall dependency length between the verb and its two arguments is shorter in V[punch] S[chef] O[referee next to blue skateboard] than in V[punch] O[referee next to blue skateboard] S[chef]. Specifically, in the VSO order, the verb and the subject are adjacent and there is only one word between the verb and the object; however, in the alternative VOS order, though the verb and object are adjacent, there are five words intervening between the verb and the subject. By contrast, for the verb-final language, the overall dependency length between the verb and its two arguments is shorter in O[blue skateboard next to referee] S[chef] V[punch] than in S[chef] O[blue skateboard next to referee] V[punch]: in the OSV language, the verb and subject are adjacent and only one word intervenes between the verb and object; however, in the SOV language, while the verb and the object are adjacent, five words intervene between the verb and the subject.
<table>
<thead>
<tr>
<th>Linguistic topic</th>
<th>Typological generalization/pattern</th>
<th>Cognitive bias in artificial language learning or communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differential case marking (DCM)&lt;sup&gt;20&lt;/sup&gt;</td>
<td>DCM exists in a number of languages (Aissen 2003; Haspelmath 2019b). The motivation for this phenomenon is debated: DCM may be driven by ambiguity avoidance (e.g., object marking for animate objects which are more likely to be confused with subjects) (Comrie 1989; Croft 2002), or by predictability-based marking (e.g., object marking for less typical objects) (Haspelmath 2019b).</td>
<td>English-speaking adults were taught a verb-final language with both SOV (60% of all sentences) and OSV (40% of all sentences) order for transitive events and optional case marking for objects or subjects (depending on experiments) regardless of their animacy. They were tested on both a comprehension and a production task. In the comprehension task, participants had higher accuracy when the animate objects or inanimate subjects were case-marked. In the production task, participants also showed a bias for conditioning case marking on animacy: inanimate subjects were more likely to be case-marked than animate subjects, and animate objects were more likely to be case-marked than inanimate objects (Fedzechkina et al. 2012).&lt;sup&gt;21&lt;/sup&gt; Smith &amp; Culbertson (2020) replicated Fedzechkina et al. (2012)’s experiments with truly ambiguous sentences. They found that English-speaking adults showed the expected DCM pattern in a communication task, but no such bias in learning (Smith &amp; Culbertson 2020).</td>
</tr>
</tbody>
</table>

---

<sup>20</sup>Differential case marking is a phenomenon in which the argument of a verb can be either marked or unmarked by morpho-syntactic means, depending on several factors, such as animacy, definiteness, or specificity of the argument. Below is an example from Spanish in which the inanimate object does not have an object marker whereas the animate object requires it (Haspelmath 2019b: 314):

(ii) a. *Vi la casa.*
    I saw the house
    ‘I saw the house.’

    b. *Vi a la mujer.*
    I saw acc the woman
    ‘I saw the woman.’

<sup>21</sup>However, as noted by Smith & Culbertson (2020), sentences in the experiments in Fedzechkina et al. (2012) are not truly ambiguous because certain animate nouns exclusively function as subjects,
<table>
<thead>
<tr>
<th>Linguistic topic</th>
<th>Typological generalization/pattern</th>
<th>Cognitive bias in artificial language learning or communication</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Tal et al. (2022) taught English-speaking adults an artificial language with equal SOV or OSV order. Subjects and objects were both animate but with different information status (i.e., a given subject and a new object, or a new subject and a given object) and an optional object marker that randomly appeared in half of the sentences. Participants were tested on both a comprehension and a production task. While no direct influence of information structure on differential object marking (DOM) was observed, an indirect effect of information structure on DOM was observed: participants were more likely to use OSV than SOV when the object was given, and OSV was more likely to have object marking.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evidentiality</th>
<th>Indirect information is more likely to be marked compared to direct information across languages (de Haan 2013; Aikhenvald 2018).</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-speaking participants were exposed to videos in which a person either witnessed an event happening, or inferred the occurrence of the event, or heard someone else describe the event. Participants were correspondingly exposed an artificial language with a marker for events with that information source. They were tested on a comprehension task and a production task. Participants exhibited higher accuracy in both tasks when learning the language with a reportative or inferential information marker compared to the language with a visual information marker (Saratsli et al. 2020).</td>
<td></td>
</tr>
</tbody>
</table>

while others exclusively function as objects. This raises the question of whether the results offer clear evidence for the ambiguity avoidance account for DCM.

22 Since given information typically appears in the subject position and new information typically appears in the object position (Du Bois 1987), sentences with a new subject and a given object are less frequent and thus less predicative than sentences with a given subject and a new object.

23 Direct information is the information for which the speaker has some sort of sensory evidence (e.g., visual evidence or auditory evidence), whereas indirect information is the information that the speaker does not directly witness but gains knowledge of afterwards. Two subclasses of indirect information are inferential information and reportative information.
<table>
<thead>
<tr>
<th>Linguistic topic</th>
<th>Typological generalization/pattern</th>
<th>Cognitive bias in artificial language learning or communication</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Indirect information is less reliable than direct information and is thus more likely to be morphologically marked (Papafragou et al. 2007; Aikhenvald 2018)</td>
<td>English-speaking participants were exposed to videos in which a person either witnessed an event happening, or heard someone else describe the event. Participants were exposed to an artificial language with a marker for events with one of the two information sources. They were tested on comprehension and production. Participants exhibited higher accuracy in both tasks when learning the language with a reportative information marker compared to the language with a visual information marker. This pattern persisted when participants were explicitly provided with the meanings of the markers, and also when the evidential meanings were conveyed using non-linguistic marking (e.g., a red frame). This suggests that the learnability of different evidential categories is influenced by the markedness of their respective meanings (Saratsli &amp; Papafragou 2023).</td>
</tr>
<tr>
<td>Personal pronoun paradigms</td>
<td>Languages in which pronouns have homophony between second and third person categories (i.e., first vs. non-first person distinction) are more common than others (Cysouw 2009: 64).</td>
<td>English-speaking adults were taught pronominal forms for two of the three person categories and tested on how they generalized these forms to the remaining untrained person category (e.g., training: forms for first person and second person; testing: which of the two forms would be used by learners to refer to the third person). Learners showed a preference for person system with homophony between second and third person over alternative systems (Maldonado &amp; Culbertson 2022).</td>
</tr>
</tbody>
</table>

---

24For example, Dutch verb agreement features a clear distinction between first-person and non-first-person forms. The singular first person is marked by a zero form, while the singular second and third person are both marked by the -t suffix (Cysouw 2009: 41).
Linguistic topic | Typological generalization/pattern | Cognitive bias in artificial language learning or communication
--- | --- | ---
Languages with first person inclusive homophony are common whereas those with second and third person inclusive homophony are rare (Cysouw 2009: 91). | English-speaking adults were taught an artificial language with either a first-inclusive, a second-inclusive, or a third-inclusive pronoun and were tested on the learning accuracy of the pronominal forms. Accuracy of learning the language featuring first-inclusive homophony and second-inclusive homophony was higher than that featuring third-inclusive homophony (Maldonado & Culbertson 2022).
<table>
<thead>
<tr>
<th>Linguistic topic</th>
<th>Typological generalization/pattern</th>
<th>Cognitive bias in artificial language learning or communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response particles for polar questions</td>
<td>A hierarchy of meanings of response particles has been hypothesized: Positive Agreement &gt; Negative Disagreement &gt; Negative Agreement &gt; Positive Disagreement (‘&gt;’ indicates less marked) (Roelofsen &amp; Farkas 2015).</td>
<td>English-speaking adults were trained on a novel three-way particle system in which two general particles were used to convey a single feature (i.e., the absolute polarity of the response or the relative polarity to previous discourse), and a third specialized particle to convey one of the four feature combinations in the hierarchy. Participants were more likely to use the specialized particle for Positive Disagreement compared to Negative Agreement, but they were best at learning the Positive Agreement particle. When English-speaking adults were exposed to an artificial language with two response particles conveying all four possible meanings in the hierarchy (both a full long form and a short chipped form were randomly used for the two particles), they tended to use the shorter form more for Positive Agreement (which is less marked), and the longer form for Positive Disagreement and Negative Agreement (which are more marked meanings) when they were taught two particles expressing Negative and Positive. However, learners showed no preference between the short and long forms when they were taught two particles expressing Agreement and Disagreement. These results go against the hypothesized connection between the hierarchy and response particle forms (Maldonado &amp; Culbertson 2023).</td>
</tr>
</tbody>
</table>

Table 1.1: Parallels between typological tendencies and cognitive biases in artificial language learning and communication.

---

25 Positive and negative refer to the absolute polarity in the response, and agreement and disagreement refer to the polarity relative to previous discourse. Here are examples of the four meanings (Maldonado & Culbertson 2023: 2):

b. A: Ally doesn’t eat meat. B: Yes (No), he does. (Positive Disagreement)
c. A: Ally eats meat. B: No, he doesn’t. (Negative Disagreement)
d. A: Ally doesn’t eat meat. B: No (Yes), he doesn’t. (Negative Agreement)
Different types of cognitive biases have been discussed to explain the behavioral preferences found in these experiments (Culbertson 2023). For example, the bias for harmonic orders in the noun phrase has been argued to be at least partly shaped by a domain-general bias for simplicity mentioned above. Similarly, the bias for homomorphic orders can also been seen as driven by a simplicity bias—word orders that transparently map between the underlying syntactic/semantic relations and linear order are simpler than those that do not (Culbertson & Kirby 2016). The trade-off between simplicity and communicative accuracy is implicated in a number of these phenomena. For example, optional plural marking, differential case marking and the asymmetric coding of meanings with different markedness are often argued to reflect this pressure for communicative efficiency (Kurumada & Grimm 2019; Smith & Culbertson 2020; Saratsli & Papafragou 2023). While the range of morphosyntactic topics studied using this approach is currently limited, the parallels between patterns observed in artificial language learning and typology suggest that learning plays an important role in shaping typological universals. To summarise, artificial language learning experiments have emerged as a promising tool for uncovering links between language structure and cognition, typically through mechanisms that are at work during learning.

26There is also some research which finds that behavioral preferences are not consistent with patterns in typology. For example, it has long been observed that animacy has an influence on the plural marking of nouns across language. Specifically, animate nouns are more likely to have plural marking than inanimate nouns (Comrie 1989). One explanation for this tendency is based on the markedness of different category meanings (Croft 2002): entities higher on the hierarchy are more salient and accessible, making them more likely to be distinguished between singular and plural forms than inanimate nouns. Chen & Schuler (2021) taught English-speaking adults and children an artificial language in which animate (i.e., farm animals) and inanimate (i.e., farm vehicles) nouns appeared with plural marking at varying frequencies. No preference for plural marking for animate nouns was observed. As discussed by the authors, the results might be partly caused by the design of the experiment: distinction between animate noun and inanimate noun in the experiment might not be salient enough (i.e., farm animals vs. farm vehicles). Failure to find evidence linking cognitive biases to typological patterns might in some cases reflect a true lack of connection, but as with all experimental work, interpreting null results is difficult.
1.5 Outline of this thesis

This thesis consists of two sets of experiments, reported in two parts. In Part I, I report three experiments testing a learning bias for cross-category harmony, one of the most well-known typological universals. Cross-category harmony was first observed by Greenberg (1963). It describes the phenomenon by which word order of some phrases tends to correlate within languages, that is, different grammatical categories tend to align with each other across phrases (e.g., verbs in verb phrases tend to align with adpositions in adpositional phrases). As introduced in Section 1.2.2, one of the most popular explanations for cross-category harmony in the literature is the Head-directionality parameter. However, the fact that not all syntactic heads in all phrases are aligned has historically cast doubt on this explanation. This has been further undermined by diachronic research which can potentially provide an alternative explanation for why certain categories align. Here I conduct three artificial language learning experiments to test the hypothesis that this universal is indeed shaped by features of our cognitive system. Specifically, I ask two questions: (i) are there learning biases for word order harmony between different types of phrases? If yes, do the learning biases match the typological tendencies we see? Specifically, I focus on word order harmony between two pairs of phrases: harmony between the verb phrase and adpositional phrase, and harmony between the verb phrase and noun phrase with an adjectival modifier. These two pairs of phrases are quite different: there is strong typological evidence for harmony between the verb phrase and adpositional phrase but no such evidence for harmony between the verb phrase and noun phrase; (ii) what factors determine why only certain phrases exhibit word order harmony?

To preview the main results, learners show a bias for harmony between the verb phrase and adpositional phrase regardless of whether the order of the two phrases in learners’ native language is harmonic (i.e., English speakers) or mixed (i.e., Mandarin speakers), but no such bias was observed between the verb phrase and noun phrase with stative adjectives, roughly mirroring typology. However, a bias for harmony between the verb phrase and noun phrase emerges when the adjectives have more active, verb-like meanings. These results suggest that the cognitive bias for
cross-category harmony is not purely based on the syntactic notions of head and dependent, but may be influenced by the semantic similarity of categories to be aligned in different phrases (i.e., verbs in the verb phrases and adjectives in noun phrases).

In Part II, I report three experiments testing learning biases for word order between numeral, (numeral) classifier, and noun. Numeral classifiers are an important feature of many languages. Typological research suggests that among the six logically possible ways of ordering these three elements, only the orders in which numeral classifiers are adjacent to numerals are attested. This has been considered one of the most important pieces of evidence distinguishing between alternative analyses of the syntactic status of classifiers: whether the hierarchical structure underlying these phrases is such that classifiers combine with numerals first, or combine with nouns first. However, typological evidence might be problematic in this case since classifiers are highly geographically constrained and likely reflect the geographic relatedness of classifier languages. Moreover, even if this typological pattern reflects the universal structure of classifiers in human language, it is still an open question what drives this. Here I conduct three artificial language learning experiments to investigate two questions: (i) are the word order patterns predicted by one of the two competing structural analyses preferred by learners? and (ii) do the learning biases match the tendency we observed in typology? Specifically, in Experiment 1, I tested whether learners have a preference between (i) Cl-N-Num order vs. N-Num-Cl order, or between (ii) Num-N-Cl order vs. Cl-Num-N order. In Experiment 2, I tested whether learners have a preference between (iii) N-Num-Cl and N-Cl-Num order. In Experiment 3, I tested whether learners have a preference between N-Num-Cl and N-Cl-Num when the classifiers are mensural rather than sortal.

However, in all experiments, learners showed no reliable preference for the word order patterns that are predicted by either of these two hypotheses. I discussed the possibility that the consistent lack of a preference across these experiments might suggest that there is no universal or preferred underlying structure for numeral classifiers (and mensural classifiers) across languages. Moreover, in all experiments, learners showed no preference for the order that is relatively more fre-
quent in typology. This disparity suggests that the cross-linguistic tendency in num-
meral classifier order may reflect the geographic relatedness of classifier languages.
Part I

Evidence for the cognitive foundations for cross-category harmony
Chapter 2

The learning bias for cross-category harmony is sensitive to semantic similarity: Evidence from artificial language learning experiments
Author contributions

This chapter consists of a manuscript resubmitted to Language in response to comments from reviewers and editors. The manuscript has been formatted to align with the overall structure of the thesis. The experiments were conceived in cooperation with my supervisors Jennifer Culbertson and Simon Kirby, who are also co-authors of the paper. I designed the experiments, collected the data, conducted the analysis and wrote the first draft of the paper. Jennifer Culbertson and Simon Kirby provided advice on the work, as well as edits and comments on the manuscript.
Abstract

Cross-category harmony is one of the most well-known typological universals. It describes a trend for word order to correlate across different types of phrases within a language. Explanations for this universal vary as to whether cognitive factors play a role, or instead the tendency is due to mechanisms of language change alone. In this paper we report a series of artificial language learning experiments that aim to test a hypothesized link between cognition and cross-category harmony. As with the typological tendency itself, we find mixed evidence for harmony across different types of phrases. Specifically, learners are biased in favor of harmony between the order of verb and object and the order of adposition and noun. However, the bias for harmony between the order of verb and object and the order of adjective and noun depends on the semantic similarity between adjectives and verbs. When adjectives are active and therefore more verb-like (e.g., ‘broken’), we find harmony; when they are stative and therefore less verb-like (e.g., ‘blue’), we do not. These results suggest that the bias for cross-category harmony is not purely based on the syntactic notions of head and dependent, but reflects the interaction between a general cognitive bias favoring consistent order and cross-category similarity.

key\textit{words}: word order universals, cross-category harmony, learning biases, simplicity, semantic similarity

2.1 Introduction

There is incredible diversity across human languages as well as variation within them. However, research in typology has shown that certain linguistic patterns recur in human languages as statistical universals, while alternative patterns rarely occur or are even unattested (e.g., Greenberg 1963). Word order harmony is one of the best-known statistical universals, describing the tendency for word order to correlate across phrases within a language. In other words, harmonic word orders occur when distinct syntactic categories align with each other across phrases. For example, languages with verb-object (VO) order tend to have preposition-noun (PrepN) order (e.g., English), while languages with object-verb (OV) order tend to
have noun-postposition (NPost) order (e.g., Japanese). This phenomenon was first noted by Greenberg (1963) (in a small sample of 30 languages). Some of Greenberg’s observations on the correlation between word orders in different phrase types are shown in (1):

(1) Greenberg’s (1963) first three cross-category harmony universals
Universal 2: In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it almost always precedes.
Universal 3: Languages with dominant VSO order are always prepositional.
Universal 4: With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional.

2.1.1 Typological evidence for cross-category harmony

Since Greenberg’s seminal work, a number of studies have attempted to provide more robust typological evidence for cross-category harmony (e.g., Hawkins 1983; Dryer 1992; Jäger & Wahle 2021). For example, Dryer (1992) investigated whether the tendencies documented by Greenberg (1963) hold for a larger sample of more than 600 languages. To control for genetic and areal relationships, languages were divided into family groups with time depths of no more than 4000 years in six large geographical areas. If the order correlations noted by Greenberg (1963) held between a given pair of phrases more commonly than the opposite order in most of these six areas, then Dryer (1992) considered the word order to be significantly correlated. Orders of sixteen pairs of elements were identified as correlated with the order of verb and object. Many, but not all of the observations made by Greenberg (1963) were confirmed based on this analysis and supported by subsequent research with more than 1500 languages for some word order pairs (see Dryer 2009).

While the stratification sampling method used in Dryer (1992) provides some confidence that these correlations are robust, it might still have some limitations. For example, the results might be biased by the particular assumptions of the genetic relations of languages in the sample (see Dunn et al. 2011; Ladd et al. 2015;
To mitigate the effects of data non-independence in language samples, Bayesian phylogenetic methods have recently been used to re-evaluate the word order correlations observed in typology. Dunn et al. (2011) used this approach to assess the evolution of eight word order features in Dryer (1992) in four large families. Dunn et al. (2011) compared two models, to see which one better fits typological data on word order changes across languages within each family. One model assumed that changes in word order across different types of phrases are correlated (e.g., if verb-object changes at some point in the history of a language, so does adposition-noun order). A second model assumed independence between changes in word order of different phrase types. They found evidence for some correlations, but only within certain language families, suggesting that this tendency may not in fact be universal for any phrase types.

However, Jäger & Wahle (2021) point out that Dunn et al. (2011) test different models for each language family, but do not actually compare models that differ in whether they assume correlations are the same across families or not. Therefore, these results cannot be used to make claims about whether patterns of word order correlation are universal in that sense. Jäger & Wahle (2021) re-assess the same eight word order features to explicitly test whether there is evidence for the same correlation patterns across languages. They used a more representative sample than Dunn et al. (2011), including data from both large families and small isolate languages (1,626 languages in total). They fitted two models for each word order pair: a universal model in which the estimated word order correlation for a pair is the same for all lineages and a lineage-specific model in which the estimated correlation for a pair is different in each lineage. They found that the universal models fit the typological data much better than the lineage-specific models in many cases. In fact, their results provided evidence for consistent correlations between thirteen pairs of phrases, largely confirming the findings of previous typological research. This suggests that cross-category harmony is indeed a robust tendency in human languages.

1The eight features involved in Dunn et al. (2011) and Jäger & Wahle (2021) are verb-object, verb-subject, adposition-noun, noun-genitive, noun-relative clause, noun-adjective, noun-numeral, and noun-demonstrative.
2.1.2 Explanations for cross-category harmony

Over the past several decades, even as new typological work has challenged or confirmed particular correlations between different types of phrases, many explanations have been put forward to account for why harmony holds and what it holds between. Indeed, harmony has played a critical role in one of the most prominent debates in linguistics and cognitive science—to what extent is language shaped by the human cognitive system? Existing explanations can be grouped into two broad classes along these lines: cognitive explanations and historical or diachronic explanations.

Cognitive explanations

Cognitive explanations generally claim that cross-category harmony is driven by linguistic or cognitive factors. A long-standing hypothesis in linguistics is that this universal reflects a high-level head-dependent rule represented in the minds of language learners and users. For example, Vennemann (1974) proposed the Natural Serialization Principle, arguing that languages serialize all their operator-operand pairs in a consistent abstract word order schema—either operator before operand, or operand before operator. A similar idea is captured in the so-called Head-direction parameter in the Principles and Parameters linguistic framework (Travis 1984; Baker 2001). According to this theory, a given language can either set this high-level parameter to head-initial or head-final. This parameter will operate across phrases, resulting in cross-category harmony. A language learner in turn can observe, for example, just one instance for SVO order, and determine the parameter setting for the language as head-initial. The Head-direction parameter was formulated as a domain-specific hard constraint on human language, and thus makes strong predictions about what types of languages can exist. Because cross-category harmony is a statistical universal, this theory has been criticized on the grounds that widespread exceptions are found, and not all types of heads across phrase types have correlated order (e.g., Aristar 1991; Dryer 1992; LaPolla 2002; Haspelmath 2008).

To account for cases of apparent non-correlation that can not naturally be ex-
plained by the notions of head and dependent, Dryer (1992) developed an alternative explanation: Branching Direction Theory. According to this theory, harmony is not based on syntactic notions of heads and dependents, but is driven by alignment of categories that share the property of being phrasal or not (i.e., consisting of a single lexical category, or a phrase). Specifically, lexical categories align with the verb (in the verb phrase) and phrasal categories align with the object (in the verb phrase). For example, the order of nouns and adjectives does not appear to correlate with the order of verb and object (Dryer 1992; Jäger & Wahle 2021). Assuming that verbs and nouns are syntactic heads (see below for additional discussion of this particular assumption), these two categories would be predicted to align. However, both adjectives and nouns in noun phrases are lexical categories, therefore neither aligns with verb. However, this theory still confronts some challenges. For example, Dryer (1992) points out that verb+object phrase order is correlated with verb+manner adverb phrase order across languages. This is unexpected, since only the object is a phrasal category, manner adverbs are not. Importantly, Dryer (1992: 124) argues that semantics may play a role here: the range of meanings conveyed by adverbs (e.g., location, time, manner, etc) is similar to that conveyed by adpositional phrases, and they might therefore pattern together. It is also notable that in many cases of harmony, there is a shared syntactic category which aligns. For example the object in a verb phrase is typically a noun, and that correlates very strongly with noun order in the adpositional phrase. Similarly, order between nouns and adjectives is correlated with the order of nouns and numerals (Dryer 2013a,d). Thus, exactly what aligns in phrases which show a tendency to correlate remains, at least to some extent, an open question.

More recently, an increasing body of research has provided evidence for the hypothesis that word order harmony is at least partly shaped by a general cognitive bias for consistency, or simplicity in language learning (Culbertson et al. 2012; Pater 2012; Culbertson & Newport 2015; Culbertson & Kirby 2016; Culbertson & Kirby 2016; Culbertson & Kirby 2016).

---

2This is not always true; Dryer (1992: 111) argues that adjectival modifiers are not fully recursive phrasal categories in many languages. However, for adjective phrases which are fully recursive, the Branching Direction Theory predicts that the order between verb phrase and noun phrase should be correlated. English provides an example: recursive adjectival phrases follow the nouns they modify, just like the recursive object follows the verb (e.g., a man eager to help others).
Newport 2017; Culbertson et al. 2020a; Culbertson & Kirby 2022). For example, Chater & Vitányi (2003) put forward an extremely general notion of simplicity as a cognitive bias: all things being equal, simpler systems are preferred by learners. While in general it is not possible to compute a universal measure of simplicity, it is nevertheless useful to approximate relative simplicity given reasonable assumptions about how learners may represent systems. In the context of language, simplicity has been defined as the length of the shortest grammatical description of the language (for discussions of the notion of relative simplicity from the perspective of information theory, see Smith et al. 2013; Kirby et al. 2015; Culbertson & Kirby 2016). In the domain of word order, it is possible to formalize two grammars, one with a single ordering rule for all phrases, and another with multiple rules for the ordering of different phrase types. In at least some grammar formalisms, the first is more systematic, has less unpredictability (or entropy), and is therefore simpler. In contrast to the Head-direction parameter, which makes predictions about possible and impossible grammars, a simplicity bias explanation for harmony makes predictions about likely and unlikely linguistic patterns (Culbertson & Kirby 2022). The prediction that harmonic grammars are easier to learn is supported by a number of recent experimental studies. For example, Culbertson et al. (2012) show that learners trained on an artificial language with variable order of nouns with adjectives and nouns with numerals regularize the variation only when the input language already tends toward harmony (i.e., when nouns tend to either precede or follow both adjectives and numerals, see also Culbertson & Newport 2015, 2017; Culbertson et al. 2020a). Culbertson & Kirby (2022) show that sequential ordering patterns instantiated in non-linguistic stimuli with distinct categories are also easier to learn when category members align across sequences—similar to harmony—than when they are not. However, not all experiments have revealed clear evidence

3This has also been pointed out by Hawkins (1982: 9): “[g]rammars with more cross-categorical generalizations will be simpler than, and hence preferred over, those with fewer.”

4For example, Christiansen & Devlin (1997: 114) proposes the Recursive Rule Interaction Constraint, arguing that “if a set of rules are mutually recursive (in the sense that they each directly call the other(s)) and do not obey head direction consistency, then this rule set will be more difficult to learn than one in which the rules obey head direction consistency.”. This is supported by his work in computational modeling, where a grammar with recursive inconsistency is more challenging to learn than one with recursive consistency.
of a bias in favor of harmony.

For example, Cook (1988) taught English-speaking children an artificial language using word-by-word glosses and whole-sentence translations. Children were trained on only a subset of possible phrase types, and then tested on the order of a novel phrase type. In some cases, children inferred a harmonic order of the novel phrase type (i.e., consistent with the order of the other phrases they had learned), but in other cases they did not. Most strikingly, learners trained on SOV and AdjN order strongly preferred a non-harmonic order of adpositions, PrepN; similarly those trained on VSO and NAdj order preferred non-harmonic NPost order. In a more recent study, Zhao & Fedzechkina (2020) exposed English-speaking adults to an artificial language with either PrepN order (prepositional language) or NPost order (postpositional language), but with equally frequent VSO and SOV orders. They found evidence for a bias for harmony between OV order and NPost order, but no bias for harmony between VO order and PrepN order.

Both the accounts of, and experiments on harmony discussed so far have focused on alignment of phrases within a language, but not necessarily within an utterance. A final explanation for cross-category harmony in terms of cognitive biases is rooted in how we process particular combinations of phrases in a single utterance. Specifically, there is a long history of research suggesting that word orders which minimize the distance between (heads and) dependents are more efficient to parse (e.g., Hawkins 1994; Futrell et al. 2015b; Hahn et al. 2020). In at least some cases, when multiple phrases appear together, non-harmonic orders involve longer dependency lengths than harmonic orders. For example, a language with harmonic verb phrase order and adpositional phrase order might involve a phrase like like \[\text{VP} \text{ball play}\] \[\text{PP} \text{with kids}\] (head-initial) or \[\text{PP} \text{kids with}\] \[\text{VP} \text{ball play}\] (head-final). These phrases involve shorter dependency lengths than phrases in a language with nonharmonic order, like \[\text{VP} \text{ball play}\] \[\text{PP} \text{kids with}\] or \[\text{PP} \text{with kids}\] \[\text{VP} \text{ball play}\]. In the latter two phrases, the two heads \text{ball} and \text{kids} are dependent on each other, but are separated by the intervening dependents \text{ball} and \text{kids}, leading to a longer dependency length. While there is extensive evidence for dependency-length minimization as a principle operative across languages, not all cases of harmony involve minimizing dependency lengths. In some cases, harmonic orders have the
same or even longer dependency length as non-harmonic orders (see also Jing et al. 2022). For example, phrases with a harmonic order of nominal modifiers like ‘two red balls’ have longer dependency length than phrases with non-harmonic order like ‘two balls red’; in the former case, the dependency between ‘two’ and ‘balls’ is longer due to the intervening adjective. Interestingly, there is a strong tendency for harmony even in this case (Dryer 2013a,d), and learners prefer such orders (e.g., Culbertson et al. 2012).

Diachronic explanations

The second class of explanations for cross-category harmony suggest that it is a by-product of diachronic processes (e.g., Bybee 1988, 2010; Aristar 1991; Mithun 2003; Anderson 2016; Collins 2019). The work discussed above by Dunn et al. (2011) is in a sense a type of diachronic explanation. The authors concluded from their models that word order correlations are family-specific, i.e., that they are features retained from a common ancestral language, not derived from cognitive biases of any kind. However, there is also another type of diachronic explanation suggesting that cross-category harmony is derived via common grammaticalisation pathways. For example, Bybee (1988: 354) argues that “one grammatical order is not established on analogy with or to harmonize with another order, rather a new grammatical construction develops in a language out of constructions that already exist and shows ordering consistent with the construction from which it developed”. In other words, the order of a pair of phrase types will be harmonic when one of the phrase types was grammaticalized from the other. A straightforward example is the correlation between the order of verb and object and the order of adposition and noun. Existing research has shown that one of the most common diachronic sources for adpositions is verbs (e.g., Givón 1979; Heine & Kuteva 2002; Dryer 2019). As order within a phrase type typically remains stable during this kind of grammaticalization process, it is natural that syntactic categories which derive from the same source should share the same order. For example, in the history of Yoruba, a verb ìfún which occurred in serial verb constructions like (2), has been grammaticalized.
into a preposition marking dative (Givón 1975: 137):

(2)  
\[ \text{mo so } fún \ o \]
\[ 1\text{sg said give 2sg} \]

‘I said to you’

Yoruba is an SVO language in which the verb precedes the object. The verb *fún* remained in the same position when it was grammaticalized into an adposition, yielding PrepN order. This order is taken to be harmonic according to Greenberg (1963) and based on the theories described above: verbs and adpositions are typically considered to be heads, and both are lexical categories that take nouns, a phrasal category, as their dependents. The grammaticalization path therefore provides a plausible explanation for the harmonic order between the two phrases. Nevertheless, it has been pointed out by several researchers that there are some word order correlations which cannot be plausibly explained by diachronic processes (Dryer 1992; Croft 2002; Dryer 2019). For example, the correlation between the order of demonstrative and noun and the order of definite marker and noun. Demonstratives are the most common grammaticalization source for definite markers, but they are actually anti-correlated with the order of verb and object (see discussion in Dryer 2019).

2.1.3 New evidence for a cross-category harmony bias from artificial language learning

The competing explanations for harmony outlined here have been in dispute for decades without a consensus. One major reason for this is that most work on cross-category harmony has been based on typological and historical data. Causal relationships between specific hypothesized mechanisms and particular language features are difficult if not impossible to establish based on such data alone. For example, typological counts of language types are shaped by a huge range of confounding factors, including genetic relations between languages, language contact, 

\(^5\text{List of abbreviations: cl: classifier, sg: singular, sfp: sentence final particle, 1: 1st person, 3: 3rd person.}\)
historical accidents, and cultural factors (see Cysouw 2005; Tily & Jaeger 2011; Ladd et al. 2015). As mentioned above, artificial language learning experiments have been used to argue in favour of a cognitive explanation for harmony, providing us with a method for testing links between human learning and recurring typological patterns (see reviews by Sinnemäki 2014; Fedzechkina et al. 2016; Culbertson 2023). However, most experimental evidence actually concerns within category harmony, and specifically harmony across different types of nominal modifiers (e.g., Culbertson et al. 2012). For harmonic patterns that truly cut across syntactic categories, the evidence is much less clear (Cook 1988; Zhao & Fedzechkina 2020).

In this paper, we use a type of artificial language paradigm which is similar in spirit to Cook (1988): participants are trained on the order of one type of phrase, and then must produce phrases of a different kind. We look to see whether the order of the phrase they were trained on has an effect on the order of the novel phrase type. The method and materials are updated to reflect more modern practice in this field (i.e., no explicit translations of phrases in the artificial language, a novel lexicon participants must learn, etc.). Importantly, as mentioned above, here we take harmony to mean consistent alignment of syntactic categories across different phrases. We remain open to different hypotheses concerning which categories align—i.e., what determines whether harmony holds between two categories. The primary question of interest in this paper is therefore whether learners in our experiments are biased in favor of particular alignment patterns. Specifically, do they align order across phrase types in a way that corresponds to observed typological tendencies for harmony? Importantly, we investigate alignment between two pairs of phrase types which have played an important role in typologically-based investigations of harmony. In the first set of experiments, we investigate whether there is a cognitive bias for harmonic alignment between the verb in a verb phrase and the adposition in an adpositional phrase. This is one of the most robust cases of cross-category harmony documented (Dryer 1992, 2013h; Jäger & Wahle 2021), and an instance where alternative theories are actively in play—a cognitive bias for consistency or simplicity, a processing bias for dependency-length minimization, and a diachronic explanation based on shared origins. In the second set of experiments, we investigate alignment between categories in the verb phrase and nominal
categories, specifically phrases consisting of an adjective and a noun. This is one of the least robust cases of alignment in the typology (Dryer 2013g), rejected as a correlation by Dryer (1992) and Jäger & Wahle (2021).

To preview, we find strong evidence of harmonic alignment between verb+object and adposition+noun. This is as predicted by cognitive accounts of harmony that are based on the idea that similar categories should align: either heads and dependents, or lexical and phrasal categories, or even alignment of identical categories, in this case nouns, which form the dependent categories in both phrase types. By contrast, we fail to find evidence for consistent alignment between verb+object and noun+adjective when the adjectives are stative adjectives. While this is surprising if learners are simply aligning nouns with nouns, it is less surprising based on other cognitive accounts described above. For example, Dryer (1992) explicitly argues that these phrase types fail to align because neither nouns nor adjectives are typically phrasal, therefore neither are driven to align with objects in the verb phrase. Similarly, while most syntactic theories describe nouns as heads (e.g., Jackendoff 1977; Valois 1991; Svenonius 1994)—predicting possible alignment with verbs—there are alternatives which treat adjectives as heads (e.g., in English: Abney 1987; Truswell 2006; in Mainland Scandinavian and German: Delsing 1993). If learners fail to consistently treat either category as a head, then there is no clear basis for harmony. However, this result allows us to explore a second goal: to better understand what makes particular categories align. In particular, we test the possibility that semantic similarity between aligning categories is important (Dryer 1992). We confirm this by showing that increasing the semantic similarity between verbs and adjectives results in a higher likelihood of alignment across these two types of phrases.

2.2 Experiment 1

In Experiment 1, we focus on harmony between the order of verb and object and the order of adposition and noun. Typological survey data show that languages tend to align verbs with adpositions and objects of the verb phrase with nouns in adpositional phrases. For example, in Dryer (2013h), these orders (VO and PrepN:
456; OV and NPost: 472) significantly outnumber the alternative orders (VO and NPost: 42; OV and PrepN: 14). We will refer to the former two orders as harmonic, as the latter as non-harmonic, following all previous literature on harmony. Since analysis of raw counts can be misleading, in this case the cross-linguistic trend for harmony has been validated by a number of other methods controlling for genetic and areal relationships between languages (Dryer 1992, 2009; Jäger & Wahle 2021). Given that this trend is robust, the question is what explains it. Here we test the hypothesis that individual learners have a bias in favour of harmony, i.e., consistent order for verbs and adpositions in these two types of phrases. We test this by training learners on the order of verb and object in a novel artificial language, and asking them to subsequently produce adpositional phrases based on their knowledge of the lexicon alone, without any explicit evidence about order in that type of phrase. In other words, they have to make an inference about what order between adposition and noun should be. The prediction from our hypothesis is that learners trained on VO order will infer PrepN order, but learners trained on OV order will infer NPost order. Crucially, alternative explanations for harmony do not make this prediction. The diachronic explanation predicts no “analogy” between phrases one way or another (Bybee 1988). An alternative cognitive explanation in terms of dependency-length minimization also does not make any predictions about inferences in the task, since participants are never required to produce or comprehend utterances in which both phrases are present together.

Importantly, we targeted two populations of adult learners in this experiment: native English speakers and native Mandarin speakers. English has harmonic VO and PrepN order. Therefore, we might expect that these participants will favour this particular pair of orders when learning a novel artificial language. Alternatively, given that they have some experience with harmony in this domain, they might favour any harmonic pairing of orders (e.g., see discussion in Culbertson et al. 2012, 2020a). We therefore sought to test an additional population with active experience of non-harmony in this domain. It is rare for languages to be strictly non-harmony, particular OV languages. However, Mandarin is considered to be a VO language, but has both prepositions denoting thematic roles and postpositions de-
noting specific locations (Liu 2003; Djamouri et al. 2013; Dryer 2013h; Paul 2014). For example:

(3) \((zai)\) men-ª(hou) kan shu

\[\text{at door behind read book}\]

‘read a book behind the door’

In (3), the object shu (‘book’) follows the verb kan (‘read’). There are two adpositions, a preposition zai (‘at’) and a postposition hou (‘behind’). When the noun within the adpositional phrase is a common noun, not referring to locations, the preposition can be omitted in some contexts, but the postposition is obligatory, as shown in (3) (for detailed analyses, see Djamouri et al. 2013; Paul 2014). When the phrase is used in isolation\(^7\) which is also the context used in our experiment (see below), as in (4), the preposition zai (‘at’) is optional, and the postposition hou (‘behind’) is required.

(4) \((zai)\) men-ª(hou)

\[\text{at door behind}\]

‘behind the door’

Therefore, Mandarin speakers have experience with both harmony (i.e, PrepN order) and disharmony (i.e., NPost order) in this domain, with obligatory postpositions potentially reinforcing the non-harmonic order. If the results of our experiment are simply due to transfer from the participants’ native language, we might expect that this would lead, if anything, to a weaker harmonic preference when Mandarin learners are trained on VO order, or a weaker bias across the board compared to English speakers. If, instead, we find a strong preference for harmony

---

\(^6\)It is also grammatical to use the combination of prepositions (denoting thematic roles) and nouns (denoting specific locations) to express this kind of meaning in Mandarin. The postpositions are grammaticalized from nouns (e.g., Liu 2003). For example, (3) can also be expressed as zai men houmian kanshu without any semantic differences. The postposition hou is grammaticalized from the noun houmian (‘back.side’).

\(^7\)The phrase can be used in isolation to answer questions like ta cang naer le? (lit., ‘he hide where srf’, ‘where did he hide?’).
regardless of native language background, and regardless of whether participants are trained on VO or OV order, then this will provide clear evidence for a cognitive bias for harmony between these two types of phrases. The predictions of a harmony bias and L1 effect for both populations in the two conditions in Experiment 1 are summarized in Table 2.1.

<table>
<thead>
<tr>
<th>Condition</th>
<th>English speakers</th>
<th>Mandarin speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>VO</td>
<td>PrepN PrepN</td>
<td>PrepN NPost (or PrepN)</td>
</tr>
<tr>
<td>OV</td>
<td>NPost PrepN</td>
<td>NPost NPost (or PrepN)</td>
</tr>
</tbody>
</table>

Table 2.1: Predicted order of adposition and noun across the two conditions of Experiment 1 based on either a harmony bias or an L1 effect for both populations. Note that if English speakers have a general preference for harmony, rather than for a particular pair of L1-like orders, they might prefer NPost in the OV condition. This highlights the importance of the Mandarin participants, who have active experience of non-harmony in this domain (in the form of obligatory postpositions).

2.2.1 Methods

The experiment, including the hypotheses, predictions and analysis plan, was pre-registered at the Open Science Framework (preregistration of Experiment 1: English speakers; Mandarin speakers).

Design

We used the extrapolation paradigm (e.g., Wilson 2006; Culbertson & Adger 2014; Maldonado & Culbertson 2022) in which learners are exposed to examples of a new language but the critical evidence for the structure of interest is held out. At the critical test, learners need to extrapolate from what they have learned to held-out data. In our experiment, learners were trained on nouns, verbs, adpositions, and the relative order of verbs and objects in a miniature artificial language (either VO or OV). In the critical test, they were asked to produce utterances with adpositions and
nouns based on the limited information they received during training. As no information about the order of adposition and noun was provided, both preposition-noun and noun-postposition orders are possible. Orders that learners produced in the testing phase will therefore indicate whether they have a preference towards harmonic order between these two phrase types.

Materials

The artificial language was comprised of four inanimate nouns, four verbs (meaning ‘kick’, ‘push’, ‘point’, and ‘punch’) and four adpositions (meaning ‘behind’, ‘in front of’, ‘left of’, and ‘right of’) (see Table 2.2 for the full lexicon). Note that some of these adpositional meanings (i.e., ‘behind’ and ‘in front of’) can be expressed by postpositions in Mandarin, and would thus be non-harmonic with Mandarin verb phrase order (VO). The nouns correspond to novel objects to discourage participants from simply translating them into words in their native language. All the lexical items are disyllabic and phonotactically possible in both English and Mandarin. For each word class, the correspondence between the artificial words and meanings was randomized for each participant.

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Adposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>fugan</td>
<td>baisu</td>
<td>nanfa</td>
</tr>
<tr>
<td>sangbi</td>
<td>kani</td>
<td>misai</td>
</tr>
<tr>
<td>pinma</td>
<td>tangga</td>
<td>mudang</td>
</tr>
<tr>
<td>nulin</td>
<td>dimu</td>
<td>leiga</td>
</tr>
</tbody>
</table>

Table 2.2: Lexicon of the artificial language as presented orthographically to participants. Pronunciations used for each word are indicated using IPA in the square brackets.

Participants were randomly assigned to one of two conditions: VO or OV. In the VO condition, they were trained on head-initial phrases consisting of the (head) verb followed by an object noun; in the OV condition, they were trained on head-final phrases consisting of an object noun followed by the (head) verb. Note that the language does not feature overt subject nouns: verb phrases were comprised only of a verb and an object, with the subject noun being omitted. As shown in
Figure 2.1, the subject performing the action was always the same boy. This same boy was also used to illustrate adposition meanings. It is worth noting that this differs from the language used in Zhao & Fedzechkina (2020), which had overt subjects, along with overt agents in adpositional phrases.\(^8\)

The language was presented to participants both auditorily and orthographically. The words were pronounced by a male native Vietnamese speaker making them sound novel to both English and Mandarin speakers.\(^9\) Sound files of verb phrases were obtained by concatenating the respective verbs and nouns. Examples of the visual stimuli were illustrated as Figure 2.1. Nouns, verbs, and verb phrases were shown as images of a novel object,\(^10\) action of a boy (e.g., punch), and the action on the object (e.g., punch the object), respectively. During training, adposition meanings were illustrated by simultaneously presenting four images in which the boy is standing at a same position relative to an object (e.g., right of). Individual such images were used to elicit specific adpositional phrases during the critical test (e.g., right of the object). For the verb phrases, the direction of the action on the object was counterbalanced (in half of the images the direction of the action was rightward, in the other half the images were flipped).

\(^8\)As noted by Zhao & Fedzechkina (2020), this may partly explain the failure to find a consistent harmony bias in their experiment. In particular, sentences with VSO order were used to train VO order, and sentences with SOV order were used to train OV order. However, these two orders do not occur equally frequently across languages. In fact, SOV is the most frequent basic order across languages, whereas VSO is a rather infrequent one (Dryer 2013e). Although their frequencies in the artificial languages were same, learners might have a preference for the typologically more frequent SOV order (see evidence from silent gesture experiments, (see evidence from silent gesture experiments in Goldin-Meadow et al. 2008; Futrell et al. 2015a). This could mask the preference for the harmonic VSO order in the prepositional language. Similarly, the inclusion of an agent in adpositional phrases in the artificial language might have affected learners’ behaviour. Learners may have treated the agent as subject, and the adposition as a verb (i.e., the whole utterance as a simple clause). If this is the case, then the results may not in fact reflect anything about harmony across these phrase types, but instead a bias for SOV order in these simple clauses.

\(^9\)The language was designed as a tone language to ensure that Mandarin speakers did not impose tones on the words themselves. There were four tone patterns in each word class: flat tone + rising tone, rising tone + falling tone, dip tone + falling tone, and flat tone + flat tone. Tone was present in the auditory stimuli, but no orthographic tone makers were used. Native English speakers likely ignored the tones.

\(^10\)The novel objects were created based on those used in Culbertson et al. (2012).
Procedure

The experiment was presented using PsychoPy (Peirce et al. 2019). Participants were instructed that they would be learning a new language and the experiment would last roughly 40 minutes. Participants were seated in a quiet testing room in front of a computer, and wore headphones through which the experimental audio was played. The experiment consisted of five phases: (i) noun training and testing, (ii) verb training and testing, (iii) verb phrase training and testing, (iv) adposition training and testing, and finally, (v) adpositional phrase testing. At the beginning of each phase, participants were instructed what they would learn next.

Phase I Noun training and testing Participants were first trained on nouns in isolation (10 trials for each of the four nouns, presented in random order). In each trial, participants were shown a picture of a novel object, and saw and heard (simultaneously) the corresponding label for the object in the new language. Participants were instructed to repeat the word after they heard it. After the exposure trials, participants were tested on what they had learned (8 trials for each of the four nouns, presented in random order). In each testing trial, participants were shown a picture of an object they had previously seen and were asked to say the word for the object. Their productions were recorded. Feedback was provided in the form of the correct response (both auditorily and orthographically) on each testing trial, regardless of whether the participant’s response was correct or incorrect.

Phase II Verb training and testing Participants were then trained on verbs in isolation. The procedure and the number of trials were both identical to the noun
training and testing phase. Participants were instructed at the beginning of the phase that in each trial they would learn a word describing what the boy is doing.

**Phase III Verb phrase training and testing** Next, participants were trained on the verb phrase (2 trials for each of the 16 verb phrases, presented in random order). Participants were instructed at the beginning of the phase that they would see a picture in which a boy is doing an action to an object and hear the corresponding utterance to describe what is happening in the picture. Procedurally, verb phrase training was otherwise identical to the noun and verb training. After the training trials, there was a matching testing phase (2 trials for each of the 16 verb phrases, presented in random order). Participants were shown four pictures in each trial. In each picture, there was a boy doing an action to an object. Participants heard an utterance that described one of the four pictures and they were asked to choose the picture that matched the description. Feedback was provided on each matching trial. The correct picture matching the description was marked with a checkmark whereas the pictures not matching the description were marked with a cross. At the same time, there was a correct sound if the participant’s answer was correct, and an incorrect sound if the participant’s answer was incorrect. The participant’s cumulative number of correct trials was displayed on the screen throughout the matching phase.

After the matching test, participants were further tested on verb phrase production (2 trials for each of the 16 verb phrases, presented in random order). In each trial, participants were shown a picture with a boy doing an action to an object that they had previously seen. Participants were asked to describe what the boy is doing. Their productions were recorded. Feedback was provided in the form of the correct response (both auditorily and orthographically) on each testing trial, regardless of whether the participant’s response was correct or incorrect.

**Phase IV Adposition training and testing** Participants were then trained on adpositions in isolation. The procedure and the number of trials were identical to the noun and verb training phase, except that in each training trial, participants were shown four pictures at a time. In all four, the boy is in a same position relative to an object, as shown in Figure 2.1. During training, participants were instructed that they would hear a word describing where the boy is in the pictures. During testing,
participants were again shown a set of four pictures, and were asked to say where the boy is in the pictures.

**Phase V Adpositional phrase testing** In the final critical test phase, participants were shown a picture showing a boy positioned relative to one of the objects (i.e., one of the pictures seen in sets during adposition training; 4 trials for each of the 16 adpositional phrases, presented in random order). Participants were instructed to describe the full picture with two words. Their productions were recorded. No feedback was provided.

**Questionnaire** After participants finished the experiment, they were asked to translate the artificial verbs and adpositions into their native language, and also list any languages that they use, along with a proficiency rating. We use these data for additional exploration and analysis below.

**Participants**

52 native English speakers and 44 native Mandarin speakers recruited from the University of Edinburgh participated in the experiment. They received £8 as compensation for completing the experiment. According to our preregistered exclusion criterion, participants whose average accuracy in any of the non-critical testing phases was below 75% were excluded from data analysis (N = 4). One participant was excluded from the analysis due to failure to follow the instruction (i.e., producing only one word at the critical testing phase). Five English participants were excluded because they reported in the post-experimental questionnaire that they were also proficient in Mandarin. These exclusions resulted in the analysis of 86 participants in total (native English speakers, VO condition: 22; native English speakers, OV condition: 21; native Mandarin speakers, VO condition: 21; native Mandarin speakers, OV condition: 22).\(^{11}\)

\(^{11}\)Experiments in this paper were approved by the Ethics Committee at the University of Edinburgh. All participants gave consent before starting the experiment.
2.2.2 Results

Recall that in the critical testing phase, participants were required to produce a novel phrase type, comprised of an adposition and a noun. In addition, there were non-critical testing phases in the experiment, testing vocabulary learning for all categories of vocabulary items, and learning of verb phrase word order. Accuracy was very high (above 95%) for all non-critical testing phases, see Appendix A.1.1 for full results and statistical models. Here we focus on the critical testing phase.

Each production in the critical testing phase was coded as either harmonic or non-harmonic, given the condition the participant was trained on. In the VO condition, PrepN order was coded as harmonic and NPost order as non-harmonic; and in the OV condition, NPost order was coded as harmonic and PrepN order as non-harmonic. 30 productions (English speakers: 5 in VO condition: 1 in OV condition; Mandarin speakers: 12 in VO condition, 12 in OV condition) were excluded from analysis as they had only a single word (e.g., just a noun, or just an adposition) or had two words but the words couldn’t be identified. Figure 2.2 shows the proportions of harmonic order produced by the two populations in the two conditions.

Following our preregistered analysis plan, we compared English- and Mandarin-speaking participants’ probability of producing harmonic order in the adpositional phrase across both conditions. We fitted a Bayesian hierarchical logistic regression model using the Stan modelling language (Carpenter et al. 2017) and the brms package (Bürkner 2017) in R (R Core Team 2018).\footnote{Note that our preregistration did not specify Bayesian regression. However, in this case standard mixed effects logistic regression models using the lme4 package (Bates et al. 2015) was problematic given the extreme regularity of participants’ responses—most participants only ever produced a single order across all trials. Using Bayesian regression helps to deal with this issue. Thus all regression models in the paper will be Bayesian. The structure of the fixed effects in each model, as well as the predictions are unchanged and reflect our preregistration.} The dependent variable was harmonic order in the adpositional phrase (1 if the order produced in the adpositional phrase is harmonic with the order in the verb phrase the participant was trained on; 0 otherwise). The independent variables were condition (VO vs. OV, sum-coded) and population (English vs. Mandarin, sum-coded), and their interaction. The model included by-participant and by-item (adposition) random intercepts (all models in this paper included these two random intercepts). Considering
Figure 2.2: Proportion of responses using harmonic order in the adpositional phrase across the two conditions and two populations in Experiment 1. Note that harmony corresponds to PrepN in the VO condition, and NPost in the OV condition. Error bars indicate 95% confidence intervals on the mean of by-participant means; colored points indicate individual participant means; the dashed line shows chance level (50%). Harmonic responses are more frequent than non-harmonic ones across the board, irrespective of condition and population.

the divergent outcomes reported in previous research, we used the default prior parameter values specified in the brms package. We ran four sampling chains for 3500 iterations with a 1750 warm-up iterations for each chain, yielding 7000 samples for each parameter in the posterior distribution. For each effect of interest, we report the expected values (in log odds) based on the model’s estimated coefficients, as well as 95% credible intervals (CIs) around each estimate.

The model posterior distribution supports the following three results. First, the preference for harmonic order in native Mandarin speakers is likely to be weaker than that in native English speakers, but the difference is not reliable ($\hat{\beta} = -2.14$, CI = [-7.46, 2.56]). Second, learners’ preference for harmonic order in the OV condition is reliably weaker than that in the VO condition ($\hat{\beta} = -6.65$, CI = [-13.03, -2.00]). Third, the difference in the preference for harmonic order in the two conditions is likely to be larger for English speakers than for Mandarin speakers, but this difference is not reliable ($\hat{\beta} = 0.80$, CI = [-4.05, 5.94]).

Based on the estimated coefficients in this model, we also calculated the esti-
mated mean probability that English- and Mandarin-speaking participants produced harmonic order in the adpositional phrase in the two conditions, as well as 95% CIs around these probabilities. The results show that both English- and Mandarin-speaking participants in both conditions tended to produce harmonic order in the adpositional phrase (English speakers, VO condition: $\hat{\beta} = 26.33$, CI = [13.95, 46.62]; English speakers, OV condition: $\hat{\beta} = 11.37$, CI = [2.92, 22.60]; Mandarin speakers, VO condition: $\hat{\beta} = 20.39$, CI = [9.76, 37.64]; Mandarin speakers, OV condition: $\hat{\beta} = 8.73$, CI = [0.76, 19.69]).

2.2.3 Discussion

In Experiment 1, we taught native English and Mandarin speakers a novel lexicon with verbs, nouns, and adpositions (in isolation), and we taught them either VO order or OV order in the verb phrase. We then tested whether they extrapolated the order in the verb phrase to the adpositional phrase. English has harmonic order between the two phrase types, PrepN and VO. Mandarin has VO order but has both (optional) prepositions and (obligatory) postpositions for the meanings tested here. Thus the order in Mandarin would be potentially fully non-harmonic, NPost and VO. We predicted that a bias for harmony would hold in both populations—mirroring the strong typological tendency for harmony between these two phrase types. However, given the differences between these two languages, we also predicted that English speakers might have a stronger bias for harmony in the VO condition, or potentially a stronger harmony bias across the board, compared to Mandarin speakers. Instead, we found that both English- and Mandarin-speaking learners tended to produce harmonic orders across both conditions: PrepN order when they were taught VO order, and NPost order when they were taught the OV order. These results point to a strong bias for harmonic order between these two phrases, independent of L1 knowledge.

Nevertheless, we did find that in both English- and Mandarin-speaking participants, the preference for harmonic order was stronger in the VO condition, i.e., when the verb phrase followed their native language order. For English speakers, this could be explained by the fact that VO and PrepN together are parallel
to English. However, it is not clear that an effect of L1 experience could explain this for Mandarin speakers, where we predicted that if anything the bias would be for NPost in this case. As mentioned above, in Mandarin, when the adpositional phrase is used in isolation as in our task, postpositions marking location information are obligatory while prepositions are optional. We predicted that, if anything, Mandarin learners might therefore have a preference for non-harmonic NPost order in the VO condition. That this was not borne out suggests again that the effect of L1 experience may not be particularly strong in this task. In principle it could be that PrepN order is actually more salient for native Mandarin speakers, since although prepositions are optional in this context, they are nevertheless more frequent in discourse.\(^{13}\) Even if that is the case, it is still surprising that the possibility of NPost order with VO did not lead Mandarin speakers to produce less harmony in our task.

Notably, all native Mandarin-speaking participants were also proficient English speakers. Therefore a final possibility is that the harmonic order of their L2 influenced their behavior. We investigated this possibility by analyzing Mandarin speakers’ proficiency ratings in English on a scale from 1 to 10 (as reported in our final questionnaire). We fitted a hierarchical Bayesian logistic regression model predicting the probability that Mandarin-speaking participants produce harmonic order in the adpositional phrase by their proficiency of English and the VP order they were trained on (both sum-coded), and their interaction.\(^{14}\) Four sampling chains for 5000 iterations with a 2500 warm-up iterations for each chain, yielding 10000 samples for each parameter in the posterior distribution. The model posterior distribution indicated no reliable effect of English proficiency on the probability of...

\(^{13}\)For example, based on the text of Mandarin translation of The Little Prince, Peng et al. (2020: 5989) shows that both the token frequency and type frequency of prepositions (token frequency: 667/933 = 71.5%; type frequency: 42/70 = 60%) are higher than those of postpositions (token frequency: 266/933 = 28.5%; type frequency: 28/70 = 40%). The preposition \textit{zai}, which is also the one used to express thematic roles in our task, is the most frequent one among all of the adpositions in the corpus (token frequency: 172/933 = 18.2%).

\(^{14}\)The model with the default prior parameter values specified in brms package did not converge. We therefore adjusted the priors to attain model convergence. We used priors centered on zero to reflect that we have no clear expectations about the parameter values. In particular, the priors we used for the intercept and the other fixed effects (including the interaction term) were all Normal(0, 5). These priors permit a wide range of values, both positive and negative.
producing harmonic order in either condition.\textsuperscript{15}

It is also worth verifying that learners acquired the adpositional meanings as intended in our experiment, and not, for example, as verbs. The latter possibility would potentially confound our interpretation of their alignment of "verbs" and "adposition" as cross-category harmony. We therefore analyzed learners’ responses to the post-experiment questionnaire, in which they were asked to translate the artificial adpositions into their native language. The summary of learners’ responses in Experiment 1 is shown in Table A.19 (for English speakers) and Table A.20 (for Mandarin speakers) in Appendix A.3. English speakers’ responses suggest that, in most cases, they translated the artificial adpositions into English adpositions.\textsuperscript{16} Mandarin speakers’ responses were also relatively consistent. They predominantly translated the artificial adpositions as postpositions (or nouns that the postpositions are grammaticalized from), or as “zai+...+Postposition/Noun” in some cases. When the phrase is used in isolation, as in our experiment, zai can function both as a verb and an adposition in Mandarin. Therefore it is not possible to differentiate between these interpretations. However, given that most other types of responses were not verbal, we take this as an indication that the meanings were treated by

\textsuperscript{15}We also analyzed the questionnaire data from English speakers to check whether experience with a second language might have had an influence on their results. 15 native English speakers in Experiment 1 reported speaking a second language with a different order from English. Word order in the verb phrase and the adpositional phrase in these languages were: OV and NPost, no dominant order in the verb phrase and PrepN, or OV and no dominant order for adpositions. We fitted a hierarchical Bayesian logistic regression model to evaluate whether harmonic order in the adposition phrase is predicted by the word order type of participant’s second language, the VP order that participant was trained on (both sum-coded), and their interaction. The model with the default priors specified in the brms package did not converge. We therefore adjusted the priors to attain model convergence. We used priors centered on zero to reflect that we have no clear expectations about the parameter values. In particular, the prior we used for the intercept was Normal(0, 10), and the priors for the other fixed effects (including the interaction term) were all Normal(0, 5). These priors permit a wide range of values, both positive and negative. Four sampling chains ran for 5000 iterations with a 2500 warm-up iterations for each chain, yielding 10000 samples for each parameter in the posterior distribution. The model posterior distribution indicated no reliable effect of second language ordering pattern on the probability of producing harmonic order in either condition.

\textsuperscript{16}We acknowledge that the English translations learners provided in the post-experimental questionnaires do not necessarily mean that learners acquired the artificial words as members of specific syntactic categories in the artificial language, as in English, since the syntactic category of words expressing the same meanings in different languages can be different and translations may not necessarily adhere to the syntax of the source language. However, the translations largely demonstrate that learners did not treat the artificial words as verbs.
learners as intended.

To summarize, the results from Experiment 1 provide clear evidence that individual learners have a cognitive bias for harmonic order between the verb phrase and the adpositional phrase—that is, they align verbs with adpositions, and object nouns with nouns in adpositional phrases. Importantly, this holds regardless of whether their native language is harmonic or mixed. This points to the possibility that the strong tendency for harmony across these two phrase types in typology is the result of a cognitive bias. However, the precise nature of the bias is less clear. As mentioned in Section 2.1.3, while harmony is often discussed as holding between syntactic heads—i.e., at quite an abstract level of analysis, harmony could also in principle be based on alignment of lexical and phrasal categories, or between elements of the same syntactic category. While all phrases in our task were lexical, participants may have used their experience with noun phrases in their language to determine how to align these categories. Moreover, in our task, the exact set of nouns was used as dependents for each type of phrase. Thus it is also possible that learners based their alignment of these dependents on the fact that they are of the same syntactic category (i.e., noun), or the same set of words. Interestingly, in principle, this same possibility holds when we consider harmony between the verb phrase and the order of adjective and noun: there is a shared category, nouns, which could potentially drive harmonic ordering. However, in the typology, the evidence for harmony between these two types of phrases is much less clear (e.g., Dryer 1992). In Experiment 2 we explore whether participants extrapolate verb phrase order to phrases with a noun and an adjective. We again share the exact same set of nouns across both types of phrases. If the cognitive bias for harmonic order—either in general, or in our task—reflects alignment based on a shared syntactic category or set of words, then we should observe similar results to Experiment 1. If instead, there is something about these two types of phrases that results in a weaker (or non-existent) tendency for harmony, then participants may not in fact extrapolate as they did in Experiment 1.
2.3 Experiment 2

In Experiment 2, we use the same design and procedure as in Experiment 1 to test extrapolation of the order of verb and object to the order of noun and adjective. Recall that in this case, different hypotheses about what drives harmony —i.e., what makes a given set of categories align—make contrasting predictions. If harmony is driven by a bias to consistently order identical syntactic categories, we might expect learners to maintain the position of nouns across these two phrase types, and therefore to align verbs with adjectives. By contrast, if harmony is based on the head-dependent distinction, then assuming nouns are dependents in the verb phrase but heads in the noun phrase, then nouns should not align with nouns. Instead, we might predict that verbs in the verb phrase should align with nouns in the noun phrase. According to Dryer (1992), the typological data suggest that there is no consistent alignment either between verbs and nouns or between verbs and adjectives (see data in Dryer 2013g). Dryer (1992) argues that this is because neither noun nor adjective are sufficiently phrasal to align with the object in a verb phrase. Given this uncertainty, here we ask whether participants exhibit any consistent pattern of alignment one way or another, either between nouns across the two phrase types (thus aligning verbs and adjectives) or between verbs and nouns. This allows us to investigate (a) whether there is a tendency to extrapolate consistently (regardless of the particular alignment pattern) across participants, as we found in Experiment 1, and (b) whether this is driven (or not) by alignment of nouns.\footnote{This does not mean that we will infer behavior consistent with a harmony bias regardless of what participants do. Rather, it means that the pattern of responses we see across participants can tell us something about the basis on which alignment might occur in these experiments, and whether this matches the patterns (i.e., lack of consistent harmonic alignment between any of the categories) that we see in typology.}

As in Experiment 1 participants were taught the order of verb and object. Here, we then taught them novel adjective meanings in isolation. At the critical test, they were asked to describe a picture with both an adjective and a noun. Here we test only English speaking participants.
2.3.1 Methods

The experiment, including the hypotheses, predictions and analysis plan, was also preregistered at the Open Science Framework (preregistration of Experiment 2).

Materials

The materials were identical to those used in Experiment 1, except that rather than adpositions, in Experiment 2 we use four color adjectives (meaning ‘red’, ‘yellow’, ‘green’, and ‘blue’). Adjectives were illustrated by four images featuring different objects in a same color (e.g., ‘blue’, see example in Figure 2.3A).

Procedure

The procedure of Experiment 2 was also identical to Experiment 1, except that the adposition training and testing phase was instead an adjective training and testing phase, and the final critical extrapolation testing phase involved producing phrases with an adjective and a noun. In the adjective training and testing phase, on each trial participants were shown a picture with four objects of the same color (see Figure 2.3A). They were instructed that they would hear a word describing the color. In the noun phrase testing phase, participants were shown a picture with a single object in one of the colors they had seen in training. They were instructed to describe the full picture using two words.
Participants

43 native English speakers recruited from the University of Edinburgh participated in the experiment and received £8 as compensation. Following our preregistration, participants whose accuracy was below 75% in any one of the non-critical testing phases were excluded from the analyses (N = 1). This resulted in the analysis of 42 participants in total (VO condition: 20, OV condition: 22). It is worth noting that English speakers have VO order and AdjN order, which involves alignment of nouns across the two phrases, but would be non-harmonic under an account of head-dependent alignment that assumes the noun as the head of a noun phrase (e.g., Jackendoff 1977; Valois 1991; Svenonius 1994). As mentioned above, we will evaluate whether either pattern is found as a consistent behavioral response in our experiment.

2.3.2 Results

Accuracy was very high (above 96%) for all non-critical testing phases (see Appendix A.1.2 for full results). Here we focus on results from the final critical testing phase, where participants had to produce a novel phrase type, comprised of a noun and an adjective.

Each production in the critical test phase was coded as either AdjN order or NAdj order. 25 productions (VO condition: 3; OV condition: 22) were excluded from analysis as they had only a single word or had two words but the words could not be identified. As noted earlier, here we ask whether learners have a preference for alignment either between verbs and adjectives or between verbs and nouns. If anything, we found that adjectives are more likely to align with verbs, and nouns with nouns across the two types of phrases. Figure 2.4 thus shows the proportion of noun phrase (NP) productions that involve a consistent placement of nouns across the two phrases, and thus alignment of verbs with adjectives: AdjN order in the VO condition and NAdj order in the OV condition. For comparison, this plot also shows the proportion of responses with harmonic alignment of the verb with the adposition produced by English speakers in each condition in Experiment 1 (i.e., VO with PrepN and OV with NPost).
Figure 2.4: Proportion of responses using harmonic order across both conditions and phrase types (PP in Experiment 1, NP in Experiment 2). Note that here what we have called ‘harmonic’ responses differ depending on the phrase type. Red points correspond to the data presented for English speakers in Figure 2.2, where harmony indicates alignment of verbs and adpositions (i.e., harmonic order: VO with PrepN, and OV with NPost). By contrast, blue points show the proportion of orders in which verbs were aligned with adjectives. We chose this somewhat arbitrarily, given the contrasting predictions outlined in the main text, however this response type was numerically higher than responses which aligned verbs and nouns. Nevertheless, while there was a strong preference for harmony between verbs and adpositions in Experiment 1, there was no evidence for consistent alignment between any of the categories in Experiment 2. Error bars indicate 95% confidence intervals on the mean of by-participant means; colored points indicate individual participant means; the dashed line shows chance level (50%).
Following our preregistered analysis plan, we compared the results of Experiment 2 with the results of English speakers in Experiment 1. As for Experiment 1, we used a Bayesian hierarchical logistic regression model with default priors. Here we estimate the probability that English speakers produce harmonic alignment between a particular set of categories across experiments and conditions. As noted above, this means extrapolation from VO to PrepN and OV to NPost for adpositional phrases in Experiment 1, and given the numerical tendency in our data, from VO to AdjN and OV to NAdj for noun phrases in Experiment 2. Both predictors were sum coded. Four sampling chains ran for 3000 iterations with a 1500 warm-up iterations for each chain, yielding 6000 samples for each parameter in the posterior distribution.

The model supports the following three results. First, the preference for harmonic alignment (of nouns) between the verb phrase and the noun phrase was reliably weaker than harmonic alignment between the verb phrase and the adpositional phrase ($\hat{\beta} = -8.55, \text{CI} = [-19.69, -1.03]$). Second, the preference for harmonic alignment between the two phrase types in the OV condition was reliably weaker than that in the VO condition ($\hat{\beta} = -9.22, \text{CI} = [-21.55, -1.51]$). Third, the difference in producing a harmonic alignment in the two conditions is likely to be larger in the adpositional phrase than in the noun phrase, but the difference was not reliable ($\hat{\beta} = 2.24, \text{CI} = [-5.78, 11.43]$).

We also calculated the estimated probability that English-speaking participants produce orders aligning nouns across the verb phase and the noun phrase in the two conditions, as well as the 95% CIs around these probabilities. The results show that participants tended to produce AdjN order (thus aligning nouns) when they were trained on the VO condition ($\hat{\beta} = 18.41, \text{CI} = [2.99, 42.01]$), but participants did not reliably produce harmonic NAdj order when they were trained on the OV condition ($\hat{\beta} = 4.47, \text{CI} = [-9.50, 19.92]$).

As in Experiment 1, we also analyzed both the influence of self-reported language experience, and learners’ responses in translating the artificial adjectives into their native language. All participants translated these words into color adjectives in English, suggesting that they were likely learning these words as adjectives. The summary of learners’ responses is shown in Table A.21 in Appendix A.3.
models analysing self-reported language experience failed to converge, however as shown in Figure A.1 in Appendix A.2, the responses of learners with different second language experience are similar across the two conditions.

2.3.3 Discussion

In Experiment 2, we taught native English speakers a novel lexicon with verbs, nouns, and adjectives (in isolation), and we taught them either VO or OV order in the verb phrase. We then tested how they extrapolated the order of the verb phrase to phrases consisting of a noun and an adjective. We compared the results of this experiment with the results of English speakers in Experiment 1, where participants were asked to extrapolate from verb phrase order to phrases consisting of a noun and an adposition. By comparing across these two experiments, we are comparing how learners treat a pair of phrases which show a very strong typological tendency for harmony—verbs with adpositions, and nouns with nouns—to a pair of phrases that show no evidence for harmony—verbs are not more likely to align with either nouns or adjectives in a noun phrase. In contrast to the clear preference for consistent alignment between verbs and adpositions (and thus nouns with nouns) in Experiment 1, participants in Experiment 2 showed no preference for consistent alignment between any of these categories. They did not consistently align verbs with adjectives (and thus nouns with noun), or verbs with nouns. Specifically, while participants in Experiment 2 were numerically more likely to produce AdjN order in the VO condition, this likely reflects learners L1 order. In the OV conditions of both Experiments, participants had to over-ride their L1 order to align nouns across phrase types. While they reliable did so in Experiment 1, they did not in Experiment 2: i.e., participants in the OV condition did not reliable produce NAdj order. In other words, across individual learners there was no consistent ordering preference influenced by the order of the verb phrase. Recall that while nouns are typically considered to be heads in the noun phrase, and thus might be expected to align with verbs rather than object nouns in a verb phrase, the typological data itself suggests that there is no word order correlation between the two phrase types (Dryer 1992). This is regardless of whether adjectives or nouns are taken as the el-
ement to align with verbs (Dryer 2013g). In this sense, our results mirror typology. Importantly, this finding also suggests that the harmony bias participants have in Experiment 1 is not driven purely by a preference for placing nouns in the same place across phrases.

Why exactly does the bias for harmony differ across these phrase types? More specifically, why did participants align nouns with nouns when the other categories to be aligned were verbs and adpositions, but not when they were verbs and adjectives? In Experiment 3, we explore the effect of similarity of these other categories. In particular, we test the intuition that adpositions are more similar to verbs in some sense than adjectives are. The active verbs we used in Experiments 1 and 2 are, as mentioned above, a common historical source for adpositions across languages (e.g., Givón 1979; Heine & Kuteva 2002; Dryer 2019). Indeed adpositions have been considered “a more or less distinctive stage on a diachronic cline from independent predicate to a grammatical marker” (Bakker & Siewierska 2002: 151). The close relationship between these two categories can also be seen in the observation that across languages, the same spatial relations are sometimes expressed by adpositions and sometimes by verbs. For example, serial verb constructions—involving one verb that indicates manner of motion and another indicating direction—are a canonical way of expressing spatial meanings in a number of languages (Aikhenvald 2006b). Verbs also share some distributional properties with adpositions: they are both complement-taking (i.e., they are phrasal in the terminology of Dryer 1992). By contrast, adjectives do not typically have this property, nor do adjectives have a diachronically consistent link with verbs, at least not the kinds of color adjectives used in Experiment 2. However, some adjectival meanings are commonly derived from verbs, for example ‘twisted’ or ‘broken’. In Experiment 3, we use

18For instance, the second verb in some serial verb constructions in Vietnamese can express a change in spatial configuration, which is roughly functionally equivalent to the preposition up in English (Hanske 2013: 197):

(i) Anh ấy lăn cái thùng lên núi
   he roll c l c. barrel ascend hill
   ‘He rolled the barrel up the hill.’
these kinds of active adjectives in order to test whether the semantic similarity between adjectives and verbs modulates the preference for harmonic ordering between them.

2.4 Experiment 3

In Experiment 3, we compare learners’ tendency to align verbs in the verb phrase with adjectives in a noun phrase across two distinct classes of adjectives: stative adjectives and active adjectives. Stative adjectives and active adjectives are two subtypes of adjectives with key differences in semantics. Stative adjectives, such as adjectives describing physical properties, have the semantic property [-activity], whereas active adjectives, such as adjectives derived from verbs, are [+activity] (Lakoff 1966; Givón 1970).\textsuperscript{19} From a semantic point of view, active adjectives are thus more similar to active verbs than stative adjectives. If a cognitive bias for cross-category harmony is sensitive to the semantic similarity between both sets of elements to be aligned—e.g., a single shared category like noun is not sufficient—then the preference for alignment between verb phrase order and noun phrase order should be stronger when the adjective in the noun phrase is active compared to when it is stative. Of course, it is worth noting that here we are explicitly treating verbs and adjectives as the potential aligning elements. Under a traditional notion of cross-category harmony this would entail the assumption that adjectives in noun phrases are heads and nouns are dependents. However here we remain open about the role of heads and dependents. Although this is the classic distinction often made in descriptions of harmony, as we have discussed above some of the most well-known research on harmony has used alternative distinctions in order to better explain typological data. Moreover, in Experiment 2, if anything, there was a qualitative trend for participants in the VO condition to use AdjN, and for participants in the OV condition to use NAdj; this suggests alignment of verbs and

\textsuperscript{19}Using a series of syntactic tests, Lakoff (1966) shows a number of differences in syntactic behaviours between the two subcategories of adjectives in English. Although the syntactic behaviours of the two sub-classes of adjectives may be different in other languages, the difference in semantics seems likely to be universal.
adjectives might occur under the right conditions. Here we are testing whether higher semantic similarity between verbs and adjectives will push participants further in this direction, setting syntactic similarity aside.\textsuperscript{20}

Importantly, in addition to the color adjectives we used in Experiment 2, we also use a different set of stative adjectives in Experiment 3: texture (or material) adjectives meaning ‘spotted’, ‘stripped’, ‘checkered’, and ‘woven’. We do this because previous research has shown that color adjectives are particularly salient (e.g., Westerbeek et al. 2015; Tarenskeen et al. 2015; Gong et al. 2016). The high salience of color could in principle have driven some participants to put the adjective first (regardless of the verb phrase order they were trained on), leading to a failure to use NAdj order in the OV condition. Using texture adjectives will allow us to replicate the findings of Experiment 2, with lower salience adjective meanings. We also build in here a replication of Experiments 1 and 2. Experiment 3 therefore manipulates, between participants, whether they extrapolated to adpositional phrases, noun phrases with stative (color) adjectives, noun phrases with stative (texture) adjectives, or noun phrases with active adjectives.

\subsection*{2.4.1 Methods}

The experiment, including the hypotheses, predictions and analysis plan, were pre-registered (preregistrations of Experiment 3: Active adjectives and color adjectives; texture adjectives).

\textsuperscript{20}It is worth noting (again) that the category status of adjectives differs across languages. We have mentioned above that some theories of nominal syntax treat adjectives as heads in particular languages (e.g., Abney 1987; Truswell 2006; Delsing 1993). In addition to that, only some languages have a distinct word class for adjectives (e.g., English). In languages without a distinct word class for adjectives, adjectival meanings are expressed by verbs (e.g., Bemba) or nouns (e.g., Hausa) (Schachter & Shopen 2010; Velupillai 2012). Even among languages with a distinct adjective class, the syntactic properties of adjectives can differ. For example, in some languages, the adjective class shares grammatical properties with verbs (verby adjectives), for instance functioning as intransitive predicates; in other languages, the adjective class shares grammatical properties with nouns (nouny adjectives), for instance functioning as copular complements (Wetzer 1992, 2013). Here we are not focused on these differences, but instead differences among adjectives based on semantics. We use typical adjectival meanings (in line with typological definitions of adjectives as "word[s] denoting a descriptive property" (Dryer 2013a)). Importantly, participants appear to learn them as we have intended.
Materials

The materials of Experiment 3 were identical to those in Experiment 2, except for the adjectives used. In Experiment 2, we used four color adjectives. Here we additionally use four texture adjectives (meaning ‘spotted’, ‘stripped’, ‘checkered’, and ‘woven’) and four active adjectives (meaning ‘broken’, ‘twisted’, ‘squashed’, and ‘flipped’). Example images of active adjectives (e.g., ‘broken’) and stative texture adjectives (e.g., ‘spotted’) used in Experiment 3 are shown in Figure 2.3B and Figure 2.3C.\(^{21}\)

Procedure

Experiment 3 was implemented using the JavaScript library jsPsych (De Leeuw 2015) and conducted on the crowdsourcing platform Prolific (see footnote 22). The procedure of the experiment was identical to that of Experiments 1 and 2 with one difference: as described in our preregistration, in a pilot of this online experiment, a rather high ratio of participants was excluded for failure to follow instructions (e.g., producing only a single word rather than a phrase in the critical testing phase). To ensure that participants paid attention to the instructions, attentional checks were used in Experiment 3. Specifically, after seeing the instructions of both the noun training phase and the critical testing phase, participants were asked to choose from two options what they would need to do in the coming phase. Participants who failed the attentional checks could not resume the experiment and were excluded from analysis.

Participants

A total of 455 self-reported native English speakers were recruited through the online platform Prolific for Experiment 3.\(^{22}\) We used the Prolific pre-screening criterion that participants self-report as monolingual native English speakers from the

---

\(^{21}\)To clearly illustrate the properties, images of objects were slightly adapted for the different adjective conditions in Experiment 3.

\(^{22}\)The series of studies reported here was interrupted by the covid-19 pandemic, therefore following Experiments 1 and 2, we switched to conducting our experiments online, using Prolific.
US and UK, aged between 18 and 40, with 100% previous task approval rate. Following our preregistration, 95 participants whose accuracy rate was below 75% in any of the non-critical testing phases were excluded from data analysis (active adjective: 33; stative texture adjective: 23; stative color adjective: 18; adposition: 21). Sixteen participants were excluded due to failure to follow the instructions in the critical testing phase, i.e., producing only a single word rather than a phrase (active adjective: 2; stative texture adjective: 7; stative color adjective: 1; adposition: 6). Twelve participants were excluded due to technical issues, i.e., no sound in their recordings or incomplete data (active adjective: 4; stative texture adjective: 5; adposition: 3). One participant was excluded for failing the attentional check. A total of 331 participants were included in the data analysis (active adjectives: 43 in VO condition, 42 in OV condition; stative texture adjectives: 42 in VO condition, 45 in OV condition; stative color adjectives: 39 in VO condition, 38 in OV condition; adpositions: 37 in VO condition, 45 in OV condition). All participants were paid the equivalent of £8.91 per hour for their participation.

2.4.2 Results

Accuracy was very high (above 93%) for all non-critical testing phases (see Appendix A.1.3 for full results). Here we focus on results from the final critical testing phase, where participants had to produce a novel phrase type.

Each production in the critical testing phase was coded as either harmonic or non-harmonic depending on the order in the verb phrase that the participant was trained on. Following our hypothesis that manipulating the similarity between the verb and the other category to which it might align—adposition or adjective—will affect extrapolation, for the VO training condition, PrepN or AdjN order are treated here as harmonic; for the OV training condition, NPost or NAdj order were treated as harmonic. 259 productions (active adjective: 79 in VO condition, 33 in OV condition; stative color adjective: 21 in VO condition, 40 in OV condition; stative texture adjective: 16 in VO condition, 22 in OV condition; adposition: 7 in VO condition, 41 in OV condition) were excluded from analysis as they had only a single word or had two words but the words could not be identified.
Following our preregistered analysis plan, we first compare the preference for harmonic order between the verb phrase and the noun phrase across the two types of stative adjectives: texture adjectives and color adjectives. We fitted a Bayesian hierarchical logistic regression model with default priors to evaluate whether the use of harmonic order in the noun phrase differs depending on the type of stative adjective. The independent variables were the subtype of stative adjective and condition (both predictors sum-coded), and their interaction. Four sampling chains ran for 3000 iterations with a 1500 warm-up iterations for each chain, yielding 6000 samples for each parameter in the posterior distribution. The results of the model indicated no reliable difference in the use of harmonic order across stative color adjectives and stative texture adjectives ($\hat{\beta} = -1.48$, CI = [-4.16, 0.93]). Following our preregistered plan, we therefore collapsed the data from the two types into a single stative adjective extrapolation phrase type for subsequent analysis.

Figure 2.5 shows the results of Experiment 3, with these two conditions collapsed.

As in Experiments 1 and 2, we fitted a Bayesian hierarchical logistic regression model with default priors to evaluate the probability that English speakers produce harmonic order across extrapolation phrase type (i.e., extrapolation to adpositional phrases, stative adjective+noun, active adjective+noun) and verb phrase training conditions. The verb phrase training predictor was sum coded; the extrapolation phrase type predictor was Helmert coded. Helmert contrast coding allows us to evaluate several different contrasts in a single model by testing each level of the extrapolation phrase type predictor against the mean of all previous levels. Here we order the levels to achieve the following comparisons: (a) active adjective+noun vs. stative adjective+noun, (b) adpositions vs. the mean of active adjective+noun and stative adjective+noun. Four sampling chains ran for 6000 iterations with a 3000 warm-up iterations for each chain, yielding 12000 samples for each parameter in the posterior distribution.

The model supports the following results.\(^{23}\) First, learners’ preference for har-

---

\(^{23}\) Additionally, we also run a Bayesian hierarchical logistic regression model with priors that are built on the results of Experiments 1 and 2. The conclusions of this model are the same to the one with default priors (see Appendix A.4 for the model details and results).
Figure 2.5: Proportion of responses using harmonic order across verb phrase training conditions and extrapolation phrase types in Experiment 3. For VO conditions PrepN and AdjN are treated as harmonic; for OV conditions NPost and NAdj are treated as harmonic. Error bars indicate 95% confidence intervals on the mean of by-participant means; colored points indicate individual participant means; the dashed line shows chance level (50%). Harmonic orders are preferred, but less so for the OV conditions, which is driven by less harmony for the stative adjectives, suggesting that semantic similarity drives harmony.
monic order in OV training conditions is reliably weaker than in VO training conditions ($\hat{\beta} = -3.58, \text{CI} = [-5.67, -1.64]$). Second, learners’ preference for harmonic order between verb phrase and noun phrase is likely to be stronger when the adjective is active compared to when the adjective is stative, but the difference is not reliable ($\hat{\beta} = 1.63, \text{CI} = [-2.59, 5.94]$). Third, learners’ preference for harmonic order between verb phrase and adpositional phrase is reliably stronger than that between verb phrase and noun phrase with adjective ($\hat{\beta} = 5.71, \text{CI} = [1.14, 10.57]$).

Comparing across verb phrase training conditions, the model also indicates that the difference in the preference for harmonic order across these conditions is larger for stative adjectives than active adjectives ($\hat{\beta} = 6.05, \text{CI} = [1.69, 10.69]$). Further comparison reveals that this interaction is driven by a reliable difference between the use of harmonic order across the two verb phrase training conditions for stative adjectives ($\hat{\beta} = -6.29, \text{CI} = [-9.26, -3.93]$; the difference between the two training conditions for active adjectives is not reliable: $\hat{\beta} = -0.93, \text{CI} = [-5.18, 2.95]$). Finally, the difference between the two training conditions for the adpositional phrase is larger than for the noun phrase, but the difference is not reliable ($\hat{\beta} = 0.65, \text{CI} = [-3.82, 5.26]$).

We further calculated the probability of producing harmonic order in each individual training condition across extrapolation phrase types. We opted to re-run the model with treatment coding, using the coefficients in the new model to calculate these probabilities. The results indicate that participants had a preference for harmony between verb phrase and adpositional phrase in both verb phrase training conditions (VO condition: $\hat{\beta} = 17.75, \text{CI} = [11.54, 24.82]$; OV condition: $\hat{\beta} = 11.27, \text{CI} = [6.22, 16.96]$). There was also a reliable preference for harmony between verb phrase and active adjective+noun in both verb phrase training conditions (VO condition: $\hat{\beta} = 10.22, \text{CI} = [5.33, 15.70]$; OV condition: $\hat{\beta} = 8.75, \text{CI} = [3.87, 14.16]$). However, a preference for harmony between verb phrase and stative adjective+noun was reliable only in the VO condition ($\hat{\beta} = 14.68, \text{CI} = [10.41, 19.77]$), not in the OV condition ($\hat{\beta} = 1.14, \text{CI} = [-2.05, 4.51]$).

As in Experiments 1 and 2, we also analyzed both the influence of self-reported language experience and learners responses in translating the artificial phrases into their native language. We fitted a Bayesian hierarchical logistic regression model.
testing whether harmonic order in the noun phrase is predicted by second language order, and verb phrase training conditions (both predictors sum-coded), and their interaction. The posterior distribution of the model indicated no reliable effects of second language experience across training conditions (see Appendix A.2 for the model details). For the adpositional phrase condition, our models analysing the second language experience failed to converge, however as shown in Figure A.2 in Appendix A.2, except for one participant, all the other participants with different second language experience in the two conditions produced harmonic order. Learners’ responses in translating the artificial phrases into their native language suggest that, in the majority of cases, they translated the artificial phrases correspondingly into noun phrase and adpositional phrases in English, indicating that they were likely perceiving the artificial phrases as noun phrases or adpositional phrases. The summaries of translations for artificial phrases of different types in Experiment 3 are shown in Table A.22 through Table A.25 in Appendix A.3.

2.4.3 Discussion

In Experiment 3, we trained native English speakers on a new language with either VO order or OV order in the verb phrase, and tested whether they extrapolated the order in the verb phrase to three different phrase types: stative adjective+noun, active adjective+noun, and adposition+noun. If semantic similarity among all elements to be aligned affects the likelihood of harmony, then we expected to see an increase in alignment of the verb and adjective when adjectives are active and more verb-like compared to when adjectives are stative and less verb-like. This was borne out in our results, mainly reflected as a difference in behavior in the OV condition for the two phrase types. As in Experiment 2, participants trained on VO order use AdjN order in both adjective conditions. However, this could reflect a preference for their L1 order. Therefore the difference found in the OV training condition is crucial here. That participants trained on OV order tend to produce NAdj order—aligning verb with adjective and noun with noun across the two types of phrases—only when adjectives are active suggests that alignment between categories is indeed sensitive to semantic similarity. As discussed in detail
above, there is evidence that adpositions are more similar to verbs than typical adjectives, and harmony between the former is very robust, while harmony between the latter is not. The consistent difference between adpositions and adjectives that we find, along with a clear effect of adjective type, bolster the claim that semantic similarity, and not just head vs. dependent or phrasal vs. non-phrasal status, may be also at play in driving cross-category harmony in typology.

2.5 General discussion

It has long been noted that word order across different phrase types tends to correlate within languages—this tendency is known as cross-category harmony. These correlations often happen in similar ways across languages, for example languages with VO order tend to have PrepN order, while languages with OV order tend to have NPost. This has led to a great many different accounts, which attempt to explain why harmony holds and characterise what causes particular syntactic categories to align. These accounts include hard constraints on syntactic representations (e.g., the Head-direction parameter as in Baker 2001), a domain-general bias for consistency or simplicity (e.g., Culbertson & Kirby 2016), dependency-length minimization (e.g., Hawkins 1994), diachronic explanations stemming from common grammaticalization pathways (e.g., Givón 1975), or accidents of history (e.g., Dunn et al. 2011). Each of these accounts often comes with a distinct theory about the notion of representational similarity that determines alignment. For example, harmony might reflect a tendency for heads and dependents to align across phrases (e.g., Travis 1984; Baker 2001), or for phrasal categories and non-phrasal categories to align (e.g., Dryer 1992), or for elements which share a diachronic source to align (e.g., Aristar 1991). Previous studies have attempted to provide evidence teasing apart these different explanations, including using typological data. Recently, Bayesian phylogenetic methods have been used to evaluate whether changes in word order across different types of phrases are correlated while controlling for confounds in typological data (e.g., Dunn et al. 2011; Jäger & Wahle 2021). However, typological data cannot tell us whether harmony is driven by any particular cognitive mechanism, or any particular notion of similarity. To adjudicate between
these, behavioral evidence is needed.

A number of researchers have used artificial language learning experiments to provide evidence linking typological patterns—including harmony—to cognitive biases (e.g., Cook 1988; Culbertson et al. 2012, 2020a; Zhao & Fedzechkina 2020). Here, we built on these previous studies, using artificial language learning experiments to test whether there is a bias for cross-category harmony between distinct types of phrases, and how this bias is affected by the similarity of the syntactic categories involved. We first tested whether a cognitive bias for harmony is found between the order of verb and object and the order of adposition and noun. This is a strong typological tendency; verbs aligning with adpositions and nouns with nouns across these phrase types. We then tested whether learners show any pattern of alignment between verbs and objects in a verb phase and nouns and adjectives in a modified noun phrase. Here, there is no typological tendency, and conflicting theories of which categories might in fact align. Finally, we varied the level of semantic similarity between the adjective and the verb, testing whether we could elicit alignment between the verb and the adjective when adjectives had stative meanings (like color or texture) compared to active meanings (like ‘twisted’ or ‘broken’). Across these studies, our results revealed a strong bias for cross-category harmony between verb phrase and adpositional phrase, but a weak (or non-existent) bias for harmony between verb phrase and noun phrase, modulated by whether adjective in the noun phrase was more verb-like.

These results differ in a number of respects from what has previously been found regarding the harmony bias in artificial language learning experiments. First, although Culbertson et al. (2012) (and subsequent studies) find a bias for harmony, it is harmony within the noun phrase—i.e., between adjectives and numerals modifying a noun—and therefore not true cross-category harmony. This is important, given that many of the alternative hypotheses discussed here target harmony of this kind, and particularly harmony between the verb phrase and the adpositional phrase. Second, our confirmation of a consistent harmony bias between these two types of phrases differs from findings reported by Cook (1988) and Zhao & Fedzechkina (2020). For example, in both of those studies, some potential evidence for harmony between the verb phrase and the adpositional phrase was found,
but it was inconsistent (e.g., found only for one verb phrase order). Here, we not only find consistent evidence for this bias, but we replicate it in speakers whose native language, Mandarin, features both prepositions and postpositions, and is therefore at least partly non-harmonic. Finally, our results provide evidence for the conditions under which alignment different syntactic categories across phrases might emerge. Here, we updated the extrapolation paradigm used by Cook (1988), and found that consistent alignment better categories is sensitive to the semantic similarity of elements to be aligned. In general, the patterns we observed in language learning to some extent reflect the patterns in typology: for harmony between verb phrase and adpositional phrase which has strong typological evidence, we observed a clear consistent learning bias in different populations; for harmony between verb phrase and noun phrase which has no typological evidence, we did not observe a reliable bias when adjectives represent stative meanings, which are among the most typical meanings of adjectives.24

In the remainder of this section, we will turn to the implications of our results for theories of the specific mechanism behind harmony.

### 2.5.1 Cross-category harmony as simplicity and similarity

A number of researchers have proposed that the bias for harmonic order is a reflex of a domain-general bias for simplicity (e.g., Culbertson & Kirby 2016, 2022). This is supported by the evidence that a harmony bias favouring consistent ordering of similar categories of elements has been found within both linguistic (for harmonic order in the noun phrase, e.g., Culbertson et al. 2012, 2020a) and nonlinguistic domains (Culbertson & Kirby 2022). In particular, Culbertson & Kirby (2016) proposed that word order harmony is a result of the interaction between a domain-general simplicity bias and some notion of similarity between linguistics categories. Although it is not necessarily straightforward to quantify simplicity, under at least

---

24Dixon (1982: 16) points out that the typical semantic types of adjectives across languages are age (e.g., ‘old’, ‘new’), dimension (e.g., ‘big’, ‘small’), color (e.g., ‘white’), human propensity (e.g., ‘kind’), value (e.g., ‘good’), psychical property (e.g., ‘soft’), and speed (e.g., ‘fast’), most of which are likely to be stative. For example, according to Givón (1970), most of the English adjectives denoting these meanings are noun-based or stative (i.e., can be described as “having QUALITY\_N”).
some representational assumptions, a grammar that has a single consistent ordering rule governing order across different phrase types is more compressible than one with multiple different ordering rules for different phrase types. However, as we have noted throughout, exactly what leads to the alignment of particular categories across different types of phrases is not totally clear. The most popular notions in the literature are syntactic head and dependent. Based on these notions, cross-category harmony is explained by the rule that heads tend to consistently precede or follow the dependent across phrases. Of course, the notion of head is not always clear. Moreover, for some phrase types, no trend for harmony has been observed regardless of which element of the phrase is treated as the head (see Dryer 1992). Both these issues are relevant for the case of noun+adjective phrases that we have tested here. If the typological data in the literature are robust, then the notion of head and dependent alone may be inadequate as an explanation of what determines similarity (i.e., alignment) of categories in cross-category harmony. Indeed, Dryer (1992) argues against head and dependent as the crucial notions, and instead argues that harmony is driven by alignment of categories that share the property of being phrasal or not (i.e., consisting of a single lexical category, or a phrase).

The results of our study point to a role for semantic similarity. In all of the phrase types we tested, nouns were a potentially-aligning category. However, this was not sufficient to lead participants to extrapolate verb phrase order to a new phrase type. Instead, we found evidence that both categories had to be sufficiently semantically similar. In particular, we found participants extrapolated the relative order of verb and object to the order of adjective and noun only when the adjective was active, and therefore more semantically similar to a verb. No such bias was found with stative adjectives (either color or texture). In other words, only when the adjective in the noun phrase was more verb-like, did a tendency to harmonize the adjective with the verb appear. We also found a robust difference in strength between the bias to align verbs with adjectives compared to adpositions—as in the typology, the latter was stronger. Given the evidence for a semantic link between adposition and verb meanings, these results could also be explained by a difference in the degree of semantic similarity. Taken together, our results provide evidence for the view that the bias for harmonic order across domains is also sensitive to the fine-grained
semantic similarity of the elements to be aligned. While the possibility has been raised in previous typological work on harmony (Dryer 1992; Croft 2002), here we provide the first evidence that beyond the notions of head and dependent and phrasal and non-phrasal category, semantic similarity also plays a role in shaping cross-category harmony.

In a more general sense, our results suggest that the cognitive bias for harmony is at its root a mechanism of analogy (or generalization). A great deal of literature in psychology and cognitive science has shown that as a core cognitive mechanism, analogical mapping is based on the similarity of target and source analogs (e.g., Blanchette & Dunbar 2000; Gentner & Kurtz 2006), and more specifically, “structural alignment highlights commonalities between two analogs” (Gentner & Maravilla 2017: 191). It is worth noting that in the context of language, semantic similarity is a component of cross-categorical similarity and may be pertinent to other dimensions, such as functional similarity. For example, our results could also be related to whether learners use different categories in the same or different ways. However, since functional similarity is likely rooted in semantic similarity in this case (e.g., that active adjectives can function as verbs but stative adjectives cannot is due to the fact that active adjectives are more semantically similar to verbs than stative adjectives), and we only manipulated the meanings of words in the experiments, we highlight the role of semantic similarity here.

2.5.2 The relevance of diachrony

As mentioned above, one prominent explanation for cross-category harmony is that it is driven by diachronic processes rather than individual-level cognitive factors (e.g., Bybee 1988, 2010; Anderson 2016; Collins 2019). In this study, we have directly tested the hypothesis that a causal link between language learning and cross-category harmony exists by conducting carefully controlled experiments. The

25For example, Greenberg (1963: 79) points out that “two basic notions, that of the dominance of a particular order over its alternative and that of harmonic and non-harmonic relations among distinct rules of order,... this latter concept is very obviously connected with the psychological concept of generalization.” This is in line with simplicity-based accounts, see also Chater & Vitányi (2003) and Culbertson & Kirby (2016).
results corroborate this causal hypothesis. However, this does not imply that diachronic processes have no role to play in shaping this typological tendency. In fact, as pointed out by a number of researchers (e.g., Croft 2002; Jacques 2013; Dryer 2019; Culbertson & Kirby 2022), the cognitive hypothesis does not necessarily contradict the diachronic hypothesis in explaining cross-category harmony. For example, Dryer (2019) suggests that a diachronic relation may augment the strength of harmony between some phrase types, leading to a particular strong word order correlation. Dryer (2019) points to harmony between the order of verb and object and the order of adposition and noun, and harmony between the order of adposition and noun and order of noun and genitive as two such cases. In each of these word order pairs, there is a common diachronic path—both verbs and head nouns in genitive phrases are two common diachronic sources for adpositions—and indeed the correlation between these two word order pairs is also stronger than most other word order correlations (Dryer 1992). This suggests that grammaticalization pathways may influence the degree of harmony in a similar way as (semantic) similarity among categories.

However, it is also worth considering whether these grammaticalization pathways are themselves influenced by cognitive factors, like category similarity. In other words, it is worth asking why some grammaticalization paths are more common than others (e.g., Berg 1998; Sinnemäki 2014; Moravcsik 2013). Indeed, both syntactic and semantic suitability (sometimes called fit) have been argued to be conditions for grammaticalization (e.g., Traugott 2010; Fischer 2011). For example, grammaticalization typically features a process of extension motivated by analogy, a mechanism dependent on similarities in form, and/or meaning between the source and target structures (e.g., Traugott 2010; Fischer 2011). It is thus possible that all things being equal, one category is more likely to evolve into another if they are more similar to each other. As discussed above, the cognitive bias for harmony which is rooted in the mechanism of analogy, is also sensitive to the similarity of categories to be aligned. The likelihood of grammaticalization—or shared diachronic source—between two categories and the strength of the harmony bias between them may thus both reflect the degree of similarity between them (Culbertson & Kirby 2022).
2.5.3 A bias, not a hard constraint

Before concluding, it is worth noting here that our results support the hypothesis that harmony results from a cognitive bias rather than a hard constraint. In our experiments, while non-harmonic orders were in some cases dispreferred by learners, they were not impossible. This is, at least on its face, contrary to the predictions made by theories of cross-category harmony which treat it as a hard constraint. For example, Biberauer et al. (2014) propose a word order constraint, the Final-over-Final Constraint, which disallows structures in which a head-initial phrase is immediately dominated by a head-final phrase. Applied to the order of verb and object and the order of adposition and noun, this constraint predicts that a grammar consisting of OV and PrepN order is impossible. However, this non-harmonic order was sometimes inferred by both English- and Mandarin-speaking participants in our experiments. The idea that harmony reflects a bias has been explicitly argued for by a number of researchers (e.g., Culbertson et al. 2013), and fits with the more general claims made in research using computational models of language evolution. Such models have shown that hard language-specific constraints on language learning, like the Final-Over-Final constraint, are not likely to co-evolve with language (Chater et al. 2009). By contrast, weak learning biases in individual learners are much more likely to arise, and can be amplified by cultural evolution, giving rise to a strong or even exceptionless universals (e.g., Kirby et al. 2007; Thompson et al. 2016). That weak biases can be amplified via cultural transmission is supported by behavioural evidence from iterated learning experiments (e.g., Kirby et al. 2008; Reali & Griffiths 2009; Smith & Wonnacott 2010; Kirby et al. 2015). If this is the case, then a weak bias for harmony may drive languages to change towards harmonic orders\(^26\), leading to the strong population-level typological trends we observe today.

\(^{26}\)Interestingly, this kind of change might be especially likely when there is variation in ordering present. For example, if a harmonic language is in contact with a non-harmonic language, then the bias predicts that all things being equal, changes to non-harmonic language is more likely to happen (see Culbertson et al. 2020a for more discussion). Note that this does not mean that changes towards non-harmonic order can never happen. Jacques (2013) documents such a change from harmonic order to non-harmonic order in Japhug Rgyalrong. Japhug Rgyalrong is an OV language, in which a coordinated clause with a motion verb and a lexical verb was grammaticalized into a single predicate. The motion verb in the clause was further grammaticalized into a prefix, leading to a non-harmonic ordering between the order of verb and object and order of affix and stem (i.e.,
2.6 Conclusion

Cross-category harmony is one of the most well-known statistical typological universals, often described as correlations between word order, or alignment of particular categories of elements, across different phrase types. Over the years, a number of hypotheses have been proposed to explain harmony, including explanations based on cognitive biases and diachronic processes. In this study, we used artificial language learning experiments to provide behavioural evidence for the link between individual-level cognition and the typological tendency for cross-category harmony. We find a bias for harmony between the order of verb and object and the order of adposition and noun, matching the strong typological tendency for harmony between these phrase types. However, we also find evidence that harmony is modulated by the similarity of the linguistic categories that must align in order to produce harmony. In particular, we find no bias for alignment between verbs in a verb phrase and stative adjectives (e.g., ‘blue’) in a modified noun phrase, despite the fact the syntactic category of nouns is present in both types of phrase. However, a weak but reliable bias emerges when the adjectives have active, more verb-like meanings (e.g., ‘broken’). These findings suggest that the cognitive bias for cross-category harmony is not purely based on the syntactic notions of head and dependent, nor is it determined purely by the presence of a shared syntactic category (e.g., noun). Rather, it is influenced by the fine-grained similarity of both categories present in a pair of phrases. We argue that this notion of similarity may also underpin the diachronic processes argued to lead to harmony, potentially unifying what have historically been treated as competing explanations.

However, as confirmed by Jäger & Wahle (2021), we expect changes toward harmony to be more likely.
Part II

Lack of evidence for the cognitive foundations for numeral classifier order
Chapter 3

The syntactic status of numeral classifiers: Evidence from artificial language learning experiments

3.1 Introduction

Almost all languages have some means to categorize objects into different classes (Aikhenvald 2006a). The two most frequent types of nominal classification systems are grammatical gender and numeral classifiers (Corbett 1991; Aikhenvald 2000; Seifart 2010). Numeral classifiers are morphemes occurring in quantifying expressions whose choice is usually based on some salient semantic properties of objects referred to by nouns (Allan 1977; Dixon 1986; Aikhenvald 2000; Kilarski & Allasonnière-Tang 2021). In numeral classifier languages, when a noun (N) is

---

1 Numeral classifiers are generally classified into two sub-classes based on their different semantic (and also some syntactic) behaviours: sortal numeral classifiers (as exemplified in (1)) and mensural classifiers (e.g., piece in three pieces of cake in English). Sortal numeral classifiers categorize objects according to the inherent properties of objects, whereas mensural classifiers provide units of measure by which the objects referred to by nouns can be quantified (Gil 2013; Aikhenvald 2000).
enumerated by a numeral (Num) or a quantifier, a numeral classifier (Cl) is usually required. Below are some examples from Mandarin:

(1) a. san *(wei) keren
   three cl guest
   ‘three guests’
b. si *(zhang) zhuozi
   four cl table
   ‘four tables’
c. mei *(liang) qiche
   each cl car
   ‘each car’

In contrast to count nouns in English which are modified by numerals directly, count nouns with numerals and quantifiers in Mandarin need to co-occur with numeral classifiers, as shown in (1). The choice of a numeral classifier is largely based on the inherent semantic properties of the object referred to by the noun. For example, the classifier *wei in (1a) is only used with respected people, the classifier *zhang in (1b) is typically used with objects with a flat surface, and the classifier *liang in (1c) is only used with vehicles. Though the semantic parameters employed in different classifier systems may reflect different cultural contexts of the respective languages, some semantic properties, such as animacy (or humanness), sex,

Unlike sortal numeral classifiers which only exist in some languages, mensural classifiers are much more common and exist in almost all languages (Greenberg 1972; Her 2017). In the literature, the term "numeral classifier" is usually used to refer to sortal numeral classifier (e.g., Gil 2013). Here we follow this tradition and use "numeral classifiers" (and sometimes also "classifiers") to refer to "sortal classifiers" unless otherwise specified.

Note that in some languages, the assignment of numeral classifiers to some nouns may be more opaque or even completely arbitrary at the synchronic level (Seifart 2018). Moreover, unlike gender systems in which a noun is typically assigned to a certain class, numeral classifiers do not typically have a one-to-one relation with nouns. In many languages, different classifiers can be used with the same noun (Dixon 1986; Aikhenvald 2000) For example, in Mandarin, *keren (‘guest’) can also appear with the general classifier *ge. In this case, the choice of classifiers depends on the context and speech style of the speaker.
physical properties (e.g., shape and size), and function, are likely to be employed by numeral classifier systems in different languages (e.g., Lyons 1977; Dixon 1986; Croft 1994; Aikhenvald 2000; Vittrant & Allassonnière-Tang 2021). For example, below is a description of the classifier system in Minangkabau:

<table>
<thead>
<tr>
<th>Semantic parameter</th>
<th>Semantic feature</th>
<th>Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Animate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Human</td>
<td>unang</td>
</tr>
<tr>
<td></td>
<td>Non-human</td>
<td>ikua</td>
</tr>
<tr>
<td></td>
<td>Flat foldable objects (e.g., paper)</td>
<td>alai</td>
</tr>
<tr>
<td></td>
<td>Round, hollow objects (e.g., rings)</td>
<td>bantuak</td>
</tr>
<tr>
<td></td>
<td>Long vertical objects (e.g., trees)</td>
<td>batang</td>
</tr>
<tr>
<td></td>
<td>Flat long thin objects (e.g., timber)</td>
<td>bilah</td>
</tr>
<tr>
<td></td>
<td>Round objects (e.g., fruit)</td>
<td>inek</td>
</tr>
<tr>
<td></td>
<td>Solid objects (e.g., houses) and abstract notions</td>
<td>buah</td>
</tr>
<tr>
<td></td>
<td>Specific classifiers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inanimate</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Numeral classifier system in Minangkabau (Aikhenvald 2000: 189)

As shown in Table 3.1, objects with different animacy and physical properties (e.g., shape, rigidity, texture) are assigned to different classifiers in Minangkabau. Unlike gender (or noun class), which is typically encoded by morphological means, numeral classifiers are frequently independent lexemes that co-occur with numerals and nouns (Aikhenvald 2000). One important issue is therefore the syntactic status of numeral classifiers, and specifically, how it is represented hierarchically. In a phrase consisting of a Numeral, a Classifier, and a Noun, e.g., [Num+Cl+N]4, does the classifier form a constituent with the numeral first, and then the noun, or does it form a constituent with the noun first then the numeral? This is still highly debated, as we discuss in detail below. In this chapter, we aim to use artificial language learning experiments to look for new evidence of the syntactic status of numeral classifiers. We first summarize the competing analyses in the literature along with the evidence for these different analyses from which we will derive the predictions of our experiments.

---

3However, numeral classifiers can be dependent forms or even affixes in some languages (for examples, see Aikhenvald 2000: 105-110)

4We use [Num+Cl+N] to represent phrases consisting of numeral, classifier and noun, abstracting away from the specific word order patterns.
3.1.1 The syntactic structure of [Num+Cl+N]: alternative analyses

Classifier-with-numeral analysis

In the literature there have been two competing analyses of the hierarchical structure of the [Num+Cl+N] phrase. One analysis argues that the classifier and the numeral form a single syntactic unit, which then combines with the noun (e.g., Greenberg 1972; Rijkhoff 1990; Croft 1994; Bale & Coon 2014) (henceforth the classifier-with-numeral analysis). The hierarchical structure (abstracting away from other details) of this analysis is shown in (2).

(2) Classifier-with-numeral analysis

One of the most important pieces of evidence for the classifier-with-numeral analysis comes from the word order of the three syntactic categories. It has been observed that numerals and classifiers tend to occur as an uninterrupted sequence across languages (Greenberg 1972). Based on a sample consisting of roughly 100 numeral classifier languages, Greenberg (1972: 28) found that among all of the six logically possible orders between numeral, classifier, and noun, only the four patterns in which the numeral and the classifier are adjacent are attested. The two word order patterns in which the noun intervenes between the numeral and the classifier are unattested in his sample. Table 3.2 shows the relative typological frequency of different word orders between the three elements observed in the literature (Greenberg 1972: 28).

Other evidence includes the morphosyntactic behaviour of numeral classifiers in certain languages. For example, evidence from constituency tests, such as the movement, substitution, coordination, and question tests, shows that in many languages, numeral and classifier form a single unit: numeral+classifier combinations can be separated from the head noun or can occur without the head noun, or can be

5Specific numbers of the different word order patterns are not provided in the literature.
used independently to answer questions (e.g., in Mandarin). Indeed, it has been argued that numeral-classifier sequences in all classifier languages can be used without the overt expression of a head noun (Greenberg 1972). By contrast, in many languages, the combination of a classifier and a noun is ungrammatical and the numeral typically requires the presence of a classifier (e.g., Mandarin, Japanese). There is also evidence from morphology and phonology which appears to support the classifier-with-numeral analysis. For instance, numerals and classifiers tend to occur as a phonological unit across languages (Greenberg 1972). In some languages, classifiers can be affixes attaching to numerals or even fused into numerals (Greenberg 1972; Aikhenvald 2000), suggesting that classifiers and numerals are likely to form a unit.

**Classifier-with-noun analysis**

An alternative analysis argues that numeral classifiers are functional heads projected by nouns, thus they combine with nouns first and then merge with numerals (e.g., Cheng & Sybesma 1999; Simpson 2008; Gebhardt 2009; Dékány 2021) (henceforth the classifier-with-noun analysis). The hierarchical structure of this analysis (again abstracting away from details) is shown in (3).

(3) Classifier-with-noun analysis

<table>
<thead>
<tr>
<th>Word order</th>
<th>Typological frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num-Cl-N</td>
<td>common</td>
</tr>
<tr>
<td>N-Num-Cl</td>
<td>common</td>
</tr>
<tr>
<td>Cl-Num-N</td>
<td>uncommon</td>
</tr>
<tr>
<td>N-Cl-Num</td>
<td>uncommon</td>
</tr>
<tr>
<td>Cl-N-Num</td>
<td>unattested</td>
</tr>
<tr>
<td>Num-N-Cl</td>
<td>unattested</td>
</tr>
</tbody>
</table>

Table 3.2: Relative typological frequency of different word orders of Num, Cl and N reported in the literature
This analysis is designed to account for the fact that nouns select for particular classifiers (Dékány 2021). Another piece of evidence supporting the classifier-with-noun analysis comes from languages in which numerals and classifiers are not always adjacent: word order patterns in which the noun intervenes between the numeral and the classifier in at least some cases have been reported in some languages. For instance, in Nung (a Tai language), Cl-N-Num order is used when the numeral is ‘one’ (Simpson 2008). Moreover, in some languages, other elements such as adjectives can be inserted between numerals and classifiers (e.g., in Mandarin, see Simpson 2008 for examples). There are also some languages in which classifiers and nouns can be used without numerals, such as Cantonese (Cheng & Sybesma 1999) and Vietnamese (Simpson & Ngo 2018). In addition to semantic selection, which occurs in all classifier languages, in some language, like Nepali, classifiers show gender agreement with nouns (Dékány 2021). This has been argued to suggest that the two are in a local relationship. Finally, although it is typologically rare, classifiers can be affixes and form a single morphological unit with nouns (e.g., Kana, Aikhenvald 2000: 111).

Both syntactic analyses are supported by distinct theories regarding the semantics of numeral classifiers. The theory supporting the classifier-with-numeral analysis argues that classifiers are required due of the semantics of numerals (Krifka 1995; Bale & Coon 2014). This theory hypothesizes that bare nouns across different languages are names of kinds, but the numeral systems are different between classifier and non-classifier languages. Specifically, numerals in non-classifier languages have an incorporated measuring function and thus can combine with count nouns directly, whereas numerals in classifier languages do not have this function and need classifiers to introduce a measuring function (Krifka 1995). By contrast, an alternative theory supports the classifier-with-noun analysis, arguing that classifiers are required due to the semantics of nouns (Chierchia 1998). Unlike nouns in the non-classifier languages which include both mass and count nouns, nouns in classifier languages are parameterized as kind-denoting arguments and have no mass/count distinction. Given that mass nouns cannot be enumerated by numerals directly, classifiers are required to convert the kinds to sets of countable individuals (Chierchia 1998).
Typological data: potential problems

As illustrated above, while both syntactic structures have been argued to represent the universal underlying structure of the [Num+Cl+N] phrase, evidence for each analysis comes from specific languages. This suggests the possibility that some languages have one structure, and some the other, i.e., depending on the specific properties of the language. In fact, it has been argued by some researchers that there is no uniform or preferred structure of the [Num+Cl+N] phrase across languages (e.g., Jenks 2010; Gil 2013; Li 2013; Bale et al. 2019; Little et al. 2022). According to this viewpoint, both of the syntactic structures mentioned above are needed—numeral classifiers in some languages or some types of numeral classifiers form a constituent with numerals, whereas others form a constituent with nouns.6

Nevertheless, typological data may point to one of the two structures as preferred: if there is a preferred underlying structure of the [Num+Cl+N] phrase in the grammar of human language, then the word order patterns that are predicted by that structure should be more common than those are not. As pointed out above, the two analyses summarized above make different predictions about which orders should be preferred: the classifier-with-numeral analysis predicts that the noun should not intervene between the numeral and the classifier, whereas the classifier-with-noun analysis predicts that the numeral should not intervene between the classifier and the noun. From this point of view, the typological data largely support the classifier-with-numeral analysis—word order patterns in which the noun intervenes between the numeral and the classifier are extremely rare.

However, as discussed in Chapter 1, typological counts are not necessarily reliable; the data represent non-independent samples (e.g., Cysouw 2005; Ladd et al. 2015). This is particularly problematic in this case because the distribution of numeral classifiers is highly geographically constrained. In a recent typological survey based on a sample of 3,338 languages, 723 languages were identified as numeral

---

6Within this camp, views also vary on the relation between the types of numeral classifiers and the two different syntactic structures. Some researchers argue that numeral classifiers in different languages may have different syntactic structures (e.g., Jenks 2010; Gil 2013; Little et al. 2022), while others argue that classifiers with different functions have different syntactic structures (e.g., Rothstein 2011; Li 2013). We will return to this in Section 3.5.
As shown in Figure 3.1, most numeral classifier languages are located in Asia. Outside of this area, classifier languages are generally rare and cluster around the Pacific Rim, spanning "three macroareas: North Asian Coast (Old World), Oceania (Pacific), and the western coastline of North America, Mesoamerica, and South America (New World)" (Nichols 1992: 200). This uneven distribution implies that typological tendencies may merely be the outcome of language contact. For example, it has been argued that numeral classifiers in Asia are likely the result of language contact (Jones 1970; Her & Li 2023). Therefore, in the case of numeral classifiers in particular, evidence beyond typology is needed to determine whether some

7It is worth noting that the high proportion of classifier languages in Asia could be due to the fact that there are more languages used in Asia in the sample. However, Her et al. (2022: 159) also calculates the proportion of classifier languages in each area and finds similar results. Proportions of classifier languages in each areas are as follows: Asia (45.1%), Americas (19.2%), Europe (8.9%), Pacific (8.8%), and Africa (3.8%).
word order patterns—and by extension underlying structures of the [Num+Cl+N] phrase—are preferred. Furthermore, even if this typological pattern reflects a universal or preferred underlying structure across languages, it is still an open question what drives this.

3.1.2 Using artificial language learning approach to detect underlying structures

Recently, artificial language learning experiments have been used to provide new sources of evidence for particular hypotheses about hierarchical structure in the nominal domain (e.g., Culbertson & Adger 2014; Culbertson et al. 2020b; Martin et al. 2019, 2020, 2024; Saldana et al. 2021; Saldana & Culbertson 2023). For example, Culbertson & Adger (2014) investigated learners’ preferences among different word order patterns of the noun phrase consisting of a demonstrative, a numeral, an adjective and a noun. The hierarchical structure of this complex noun phrase has been argued to involve adjectives forming a constituent with nouns first, then numerals and then demonstratives. Word orders that reflect this, i.e., with adjectives ordered linearly closest to nouns, and demonstratives farthest away are also more common in typology. Culbertson & Adger (2014) found that when there was no explicit evidence on the word order between these modifiers, learners overwhelmingly tended to produce orders with adjectives closest to nouns and demonstratives farthest away, conforming to the hypothesised underlying structure. These results have been replicated across modalities in gesture experiments (Culbertson et al. 2020b) and in artificial language learning experiments across different populations, including populations whose native language doesn’t conform to the typical ordering pattern (Martin et al. 2019, 2020, 2024).

There is also evidence for learning biases favouring ordering of different types of nominal morphemes as well (Saldana et al. 2021; Saldana & Culbertson 2023). For example, across languages, when distinct case and number morphemes both precede or both follow the noun base, the number morpheme almost always comes between the noun base and the case morpheme (Greenberg 1963). One explanation is that this linear order again reflects the hierarchical structure. It has been argued
that morphemes that more directly modify the semantic content of the base have narrower scope and thus combine semantically with the stem first (Baker 1985; Bybee 1985; Rice 2000, see reviews in Saldana et al. 2021). Number morphemes arguably more directly modify objects referred to by noun stems, whereas case morphemes indicate an external relation between objects and events. Number morphemes therefore have narrower scope than case morphemes and should be placed closer to noun bases. As with the typological tendency itself, Saldana et al. (2021) found that when there is no explicit evidence in the input, learners strongly preferred to order the number closer to the noun stem than case. A similar pattern for the order between number, gender, and noun base was also observed in artificial language learning: learners strongly prefer to order gender morphemes closer to the noun stem than the number morpheme. This aligns with theories which argue that gender is represented together with or closer to the noun than number (Harris 1991; Alexiadou 2004; Kramer 2016, see reviews in Saldana & Culbertson 2023).

Here, we extend this line of research and take an artificial language learning approach to provide evidence for the hierarchical structure of [Num+Cl+N] phrases. If there is a preferred underlying structure of this phrase, then we should see a preference for word order patterns mapping that structure. Specifically, if the classifier combines with the noun to the exclusion of the numeral, then we should see a preference for orders in which the classifier and the noun are adjacent. By contrast, if the classifier combines with numeral to the exclusion of the noun, then we should see a preference for orders in which the classifier and the numeral are adjacent.

3.2 Experiment 1

3.2.1 Experiment 1A

Methods

Experiment 1A, including the hypotheses, predictions and analysis plan, was pre-registered here.
Design

The experiment used the “poverty-of-the-stimulus” or extrapolation paradigm (e.g., Wilson 2006; Culbertson & Adger 2014; Martin et al. 2020). Learners were exposed to a new miniature artificial language consisting of nouns, numerals, and classifiers, but no information regarding the word order between any two of the three syntactic categories was provided during training. At the critical test, learners were asked to choose, from two possible options, the most likely order of the three syntactic categories in the language. Participants’ selections will indicate whether they have a preference for certain word order patterns. It is important to note that this design differs from previous experiments on artificial language learning of word order using this paradigm (e.g., Culbertson & Adger 2014; Martin et al. 2020). For example, in previous studies on order in complex noun phrases, learners were taught the order between the noun and one modifier and tested on the order between the noun and two modifiers. By contrast here, no information about order of any of the elements is provided. This design choice is based on the particular syntactic and semantic properties of numeral classifiers. First, numeral classifier languages usually require a classifier in numeral constructions. If learners are taught "Num+N" without the classifier, this might encourage an unusual analysis of the classifier. Similarly, training learners on "Cl+N" combination without a numeral is also potentially problematic. In languages where Cl+N combinations are grammatical, this construction usually refers to objects with certain referential meanings (e.g., definite (e.g., Cantonese, Vietnamese), indefinite (e.g., Vietnamese), and/or generic (e.g., Cantonese)). The relation between the function of noun classification and reference-denoting might not be straightforward for learners, particularly within a miniature artificial language. For these reasons, we therefore taught participants lexical items, but no syntax, i.e., no order. Of course, an alternative would be provided more syntactic evidence, i.e., show participants that numerals cannot occur without classifiers, or that classifier+noun constructions have particular meanings. But these are not universal features—they do not hold across all languages—and including them would likely push participants to infer a particular structure. Here we are interested in participants’ inferences in the absence of this kind of additional
evidence. We return to these issues in the discussion section below.

Experiment 1A has two conditions, which differ based on the choices participants are given at test. In Condition 1 learners must choose between Cl-N-Num order and N-Num-Cl order; in Condition 2 learners must choose between Num-N-Cl order and Cl-Num-N order. The two structural hypotheses make different predictions regarding which word order patterns should be preferred by learners in the two conditions. Specifically, the classifier-with-noun hypothesis predicts that learners will have a preference for Cl-N-Num order in Condition 1 and a preference for Num-N-Cl order in Condition 2. By contrast, the classifier-with-numeral hypothesis predicts that learners will have a preference for N-Num-Cl order in Condition 1 and a preference for Cl-Num-N order in Condition 2. Predictions of the two hypotheses are summarized in Table 3.3.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Word order</th>
<th>Prediction Classifier-with-noun</th>
<th>Prediction Classifier-with-numeral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition 1</td>
<td>Cl-N-Num</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>N-Num-Cl</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Condition 2</td>
<td>Num-N-Cl</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>Cl-Num-N</td>
<td>×</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 3.3: Orders predicted based on the two competing hypotheses of interest in Experiment 1A.

**Material** The miniature artificial language used here is very simple, including just two lexical items in each syntactic category: two nouns (‘flower’, ‘teenager’), two numerals (‘two’, ‘three’) and two sortal numeral classifiers (one for human and one for non-human) (see Table 3.4 for the full lexicon). We employed a human/non-human distinction in our numeral classifier system since this two-way distinction is frequently observed across languages (Aikhenvald 2000: 286). In fact, animacy and humanness are the semantic features that are most likely to be distinguished in numeral classifier systems across languages (Croft 1994).

All of the artificial words have a VCVC structure. In order to minimize the confusibility of artificial words across syntactic categories, we prioritized intra-
category similarity (following Martin et al. 2020). Artificial words within the same category only differ in the first consonant. The three artificial word sets were randomly assigned to the three syntactic categories for each participant.

<table>
<thead>
<tr>
<th>Syntactic category 1</th>
<th>Syntactic category 2</th>
<th>Syntactic category 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>rovo</td>
<td>maha</td>
<td>dugu</td>
</tr>
<tr>
<td>zovo</td>
<td>naha</td>
<td>pugu</td>
</tr>
</tbody>
</table>

Table 3.4: Lexicon of the miniature artificial language in Experiment 1A. Sets of similar lexical items were randomly assigned to a category (nouns, numerals, classifiers) for each participant.

The language was presented to participants orthographically. Examples of the visual stimuli are illustrated in Figure 3.2.\textsuperscript{8} Noun meanings were illustrated by images of single objects, and numeral meanings were illustrated with a specific number of the same objects. Classifier meanings were illustrated by simultaneously presenting four images in which the objects share the same human or non-human property.

**Procedure** The experiment was implemented using the JavaScript library jsPsych (De Leeuw 2015) and conducted on the crowdsourcing platform Prolific.

\textsuperscript{8}The image stimuli used in Experiments 1A, 1B, and 2 are adapted from Vihman et al. (2018).
Participants were instructed that they would be learning a new language named Bahareti, which is spoken by roughly 50,000 people in the Southeast Asia and has some different features from English. Their participation would help us to understand how native English speakers learn and use this language. The experiment lasted roughly 10 minutes and consisted of five phases: (i) noun training and testing, (ii) numeral training and testing, (iii) classifier training and testing, (iv) mixed testing for all words across the three syntactic categories, and finally, (v) phrase testing. At the beginning of each phase, participants were instructed what they would need to do next.

Participants were first trained on nouns in isolation (6 trials for each of the two nouns, presented in random order, with no more than two consecutive trials for the same noun). In each trial, participants were shown an image of an object, and saw the label for the object in the new language. They were instructed to try to learn the labels. After the exposure trials, participants were tested on what they had learned (6 trials for each of the two nouns, also presented in random order, with no more than two consecutive trials for the same noun). In each testing trial, participants were shown an image of an object they had previously seen with two buttons under the image, each having a label. Participants were instructed to click on the label for the object in the image. Feedback was provided on each testing trial. The correct label for the object was marked with a tick and the wrong one with a cross. Participants’ cumulative number of correct trials was displayed at the top right corner of the screen throughout the testing phase.

Participants were then trained and tested on numerals in isolation. The procedure and the number of trials were identical to the noun training and testing phase. Participants were instructed at the beginning of the phase that in each of the following trials, they would see an image of a certain number of objects and also a word to express the number in Bahareti.

Next, participants were trained and tested on classifiers in isolation. The procedure and the number of trials were identical to the noun and numeral training and testing phase, except that in each trial, participants were shown four images at a time, as shown in Figure 3.2. Participants were instructed that they would learn a special set of words that are needed for counting in Bahareti. They were told that
the special word for different objects may be different. The objects in the four images in each trial all use the same special word. During testing, participants were again shown a set of the same four images they had previously seen, and were asked to choose the special word for counting the objects in the images.

Participants were subsequently tested on all of the words they had learned (6 trials for each of the 6 words, presented in random order, with no more than two consecutive trials for the same word). The testing trial of each word was identical to that in the previous testing phase.

Participants were lastly tested on the phrase consisting of a noun, a numeral, and a classifier, using a forced-choice task (3 trials for each of the four images, presented in random order). In each trial, participants were shown an image of a certain number of objects and two possible descriptions below the image. The two descriptions always included the correct lexical items but differed in the word order. In Condition 1, the two descriptions were either with Cl-N-Num order or with N-Num-Cl order. In Condition 2, the two descriptions were either with Num-N-Cl order or with Cl-Num-N order. The positions of the descriptions with the two different orders were randomized for each trial. Participants were instructed to click on the description that they believed native Bahareti speakers would use to describe the image. No feedback was provided.

After finishing the experiment, participants were instructed to complete a questionnaire asking them to translate the artificial words into their native language, their language experience, and whether they had used any strategy to choose a phrase in the last phase.

Participants

73 native English speakers recruited through the online platform Prolific participated in Experiment 1A. To ensure that participant were monolingual native English speakers, we used the following pre-screening criteria on Prolific: 1) First language: English; 2) English speaking monolingual: I only know English; 3) Were you raised monolingual? I was raised with my native language only; 4) Bilingual: none just my native language. Following our preregistration, 13 participants whose
accuracy rate was below 85% in at least one of the non-critical testing phases were excluded from data analysis (Condition 1: 6; Condition 2: 7). No participants reported speaking a language with numeral classifiers. 60 participants were included in the data analysis (30 for each condition). All participants were paid the equivalent of £9.51 per hour for their participation.

Results

The vocabulary accuracy was very high (above 97%) for all of the three syntactic categories in the non-critical testing phases, see Figure B.1 in Appendix B.1.1 for the results. Here we focus on the critical testing phase in which participants were asked to choose from two phrases with different orders.

Each selection in the critical testing phase was coded as either predicted by the classifier-with-noun hypothesis or not (1 if the order is predicted by the classifier-with-noun hypothesis; 0 otherwise. For Condition 1, Cl-N-Num order was coded as 1 and N-Num-Cl order as 0; For Condition 2, Num-N-Cl order was coded as 1 and Cl-Num-N order as 0.) Figure 3.3 shows the proportions of selecting the order predicted by the classifier-with-noun hypothesis in the two conditions.

Following our preregistered analysis plan, to determine whether the likelihood of learners choosing the orders predicted by the classifier-with-noun hypothesis is reliably above chance level in each condition, we fitted a Bayesian hierarchical logistic regression model using the Stan modelling language (Carpenter et al. 2017) and the brms package (Bürkner 2017) in R (R Core Team 2018). The independent predictor variable was Condition (sum-coded). By-participant and by-item (image in the critical testing phase) random intercepts and by-item random slopes for Condition were included. We used the default prior parameter values specified in the brms package.

The estimated coefficient of Condition in the model indicates that learners in

---

9In our preregistration, we did not specify Bayesian regression. However, the standard mixed effects logistic regression models using the lme4 package (Bates et al. 2015) is problematic because the random effects that are appropriate for our data result in fitting issues. Using Bayesian regression resolves this. Thus all regression models in this chapter will be Bayesian. The structure of the models and the predictions are unchanged and reflect our preregistration.

10This applies to all of the models reported in this chapter.
Figure 3.3: Average proportion choice of orders predicted by the classifier-with-noun hypothesis across the two conditions (i.e., Cl-N-Num in Condition 1 or Num-N-Cl in Condition 2) in Experiment 1A; error bars indicate 95% confidence intervals on the mean of by-participant means; pink points indicate individual participant means; the dashed line shows chance level (50%). Learner showed no preference for either order in Condition 1 and a weak preference for Num-N-Cl order in Condition 2.
Condition 1 were less likely to select the order predicted by the classifier-with-noun hypothesis than in Condition 2, but the difference is not reliable in the 95% credible intervals ($\hat{\beta} = -1.11, 95\%\text{CI} = [-2.47, 0.15]$). Specifically, learners in Condition 1 showed no preference for either order ($\hat{\beta} = -0.57, 95\%\text{CI} = [-2.33, 1.15]$); learners in Condition 2 were likely to choose the order predicted by the classifier-with-noun hypothesis, but the preference is not reliable in 95% credible intervals ($\hat{\beta} = 1.64, 95\%\text{CI} = [-0.17, 3.60]$). However, within the narrower 90% credible interval, the preference for learners in Condition 2 to choose the order predicted by the classifier-with-noun hypothesis becomes reliable ($\hat{\beta} = 1.64, 90\%\text{CI} = [0.13, 3.21]$).

Note that our miniature artificial language has two classifiers: one is for human and the other for non-human. In our particular stimuli at least, objects in the category of non-human are more diverse than those in the category of human. Therefore we checked to see whether word order preferences for the two classifiers differed. We ran a Bayesian hierarchical logistic regression model predicting whether learners selected the order predicted by the classifier-with-noun hypothesis based on both Condition and Classifier (both sum-coded). By-participant and by-item (noun in the phrase) random intercepts, by-participant random slopes for Classifier, and by-item random slopes for Condition were included. The estimated coefficient of Classifier in the model indicates that there is no reliable difference between the word order preferences for the two classifiers ($\hat{\beta} = -0.15, 95\%\text{CI} = [-1.11, 0.74]$).

Discussion

In Experiment 1A, we taught native English speakers a miniature artificial language consisting of two nouns, two numerals, and two classifiers, and tested whether they have a preference when they must choose between (i) Cl-N-Num order and N-Num-Cl order (Condition 1); or (ii) Num-N-Cl order and Cl-Num-N order (Condition 2). We found a weak preference for the order predicted by the classifier-with-noun hypothesis in Condition 2, but no preference for either order in Condition 1.

Before interpreting the results, it is worth verifying that learners did acquire the meanings of the artificial words, particularly the classifiers. We examined learn-
ers’ responses in the translation task in the post-experimental questionnaire. The summary of learners’ responses in Experiment 1A is shown in Table B.1 in Appendix B.2. Responses suggest that most learners did correctly learn the artificial language.

There are several possible explanations for the mixed results in Experiment 1A. One possibility is that the task might be too difficult for learners as they were provided no syntactic or distributional information at all regarding the three syntactic categories in the experiment. While we followed previous work in hypothesizing that the meanings of the categories alone might lead to ordering preferences (e.g., Culbertson & Adger 2014; Saldana et al. 2021; Saldana & Culbertson 2023), in this case there are a number of additional factors that might influence learners’ choices. For example, they might base their choice on a preference for the position of a particular category (e.g., noun in the first position or numeral in the first position). To better understand the factors that might influence learners’ choices, we analyzed learners’ responses in the questionnaire regarding whether they used any strategy in the last test. The counts of different strategies that learners reported are summarized in Table 3.5.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Condition 1</td>
</tr>
<tr>
<td>Position of a particular syntactic category</td>
<td>6</td>
</tr>
<tr>
<td>English order</td>
<td>1</td>
</tr>
<tr>
<td>Block order</td>
<td>1</td>
</tr>
<tr>
<td>No strategy</td>
<td>13</td>
</tr>
<tr>
<td>Other(^{11})</td>
<td>2</td>
</tr>
<tr>
<td>Irrelevant response (^{12})</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 3.5: Counts of different strategies reported by participants in Experiment 1A

As shown in Table 3.5, 16 participants made their choices purely based on the position of one particular syntactic category. Since the categories that learners fo-

\(^{11}\)Strategies describing the specific order without reporting the reason for choosing the order were coded as “Other” (e.g., “I went with a grammar of: type of object number of objects identity of object.”).

\(^{12}\)Irrelevant responses include responses that are irrelevant to the question (e.g., “just memory at this stage”).
cused on vary and the positions of all syntactic categories in the test orders are completely different, this could contribute to the mixed findings. In fact, 7 participants in Condition 2 reported that they chose the order with number first (i.e., Num-N-Cl), in which the order of Num and N is also the same as that in English—for example if participants neglected the syntactic category of classifiers.

Another possibility is that the mixed results might be caused by the type of task used in our experiment. In Experiment 1A, participants were asked to choose between two possible orders. Previous research has shown that compared to production tasks, perception tasks are less likely to encourage learners to create a mental representation of structures (Fawcett 2013; Martin et al. 2020). For example, Martin et al. (2020) taught English speakers the combination of a noun and a modifier (i.e., adjective, numeral, or demonstrative) in a new language and tested them on the combinations of noun and two modifiers. In a production task, they found that English speakers exhibited a bias for the order mapping the underlying structure, but this bias was not observed in a forced-choice task. As discussed in Martin et al. (2020), this might be because learners did not necessarily fully access the lexical items or form a representation of the meaning or structure of the phrase when they only needed to make a choice. If this is the case, we might expect to see a preference emerge in a production task.

Finally, there is also a possibility that our mixed results reflect that there is a lack of universal or preferred structure of the [Num+Cl+N] phrase, as introduced earlier. If this is the case, then we would not in fact expect a consistent cognitive bias favoring grouping classifiers with nouns or with numerals.

To narrow down these different possibilities, we first replicate Experiment 1A with a different condition in which the position of the noun is kept the same: in this condition, participants must choose between N-Num-Cl order and N-Cl-Num order. By having these two orders at test, learners are provided with part of the grammatical order in this new language: both the numeral and classifier follow the noun. This will hopefully reduce the variance we likely found in Experiment 1A from participants somewhat idiosyncratically wanting to put particular categories first (e.g., nouns or numerals). However, it crucially still leaves the relative order of numeral and classifier ambiguous.
3.2.2 Experiment 1B

Methods

Experiment 1B, including the hypotheses, predictions and analysis plan, was pre-registered here.

Design

The design of Experiment 1B is the same as Experiment 1A, except that in Experiment 1B there is only one condition targeting two word orders: N-Num-Cl order and N-Cl-Num order. The two hypotheses make different predictions about the two orders: the classifier-with-noun hypothesis predicts that learners have a preference for N-Cl-Num order; the classifier-with-numeral hypothesis, however, does not predict a preference for one order over another as both orders are possible under this analysis. Predictions of the two hypotheses are summarized in Table 3.6. If there is a preference for the N-Num-Cl order in the results, then it will provide evidence for the classifier-with-numeral hypothesis. However, a preference for the N-Cl-Num order will not be able to provide evidence to tease apart the two hypotheses.\(^{13}\)

<table>
<thead>
<tr>
<th>Word order</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-Num-Cl</td>
<td>×</td>
</tr>
<tr>
<td>N-Cl-Num</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 3.6: Orders predicted based on the two competing hypotheses of interest in Experiment 1B.

The material and procedure of Experiment 1B are the same as Experiment 1A.

\(^{13}\)We recognize that this makes the interpretation of results complex, but it is inevitable when the noun or the numeral is placed either at the first or the last position of the phrase. One possible design is ordering the classifier at the first or the last position and asking learners to complete the phrase with a numeral and a noun (i.e., Cl at the first position: Cl-Num-N and Cl-N-Num, or classifier at the last position: N-Num-Cl or Num-N-Cl). Under this design, each word order is only predicted by one hypothesis. However, in this case, there is a concern that learners would be likely to simply ignore the classifier and use the order of numeral and noun from their native language (i.e., Num-N).
Participants

35 native English speakers recruited through the online platform Prolific participated in Experiment 1B. The pre-screening criteria, data exclusion criteria and the payment for participants of Experiment 1B are identical to Experiment 1A. 5 participants were excluded from data analysis due to the failure to meet the accuracy criteria. No participants in Experiment 1B reported speaking a language with numeral classifiers. 30 participants were included in the data analysis.

Results

Vocabulary accuracy was very high (above 98%) for all three syntactic categories in Experiment 1B. See Figure B.2 in Appendix B.1.1 for the results.

In contrast to Experiment 1A, each selection in the critical testing phase in Experiment 1B was coded as either only predicted by the classifier-with-numeral hypothesis or not (1 if learners selected the N-Num-Cl order; 0 otherwise). Figure 3.4 shows the proportions of selecting the N-Num-Cl order in Experiment 1B.

We fitted a Bayesian hierarchical logistic regression model to determine whether the likelihood of learners choosing N-Num-Cl order was reliably above chance level. By-participant and by-item (image in the critical testing phase) random intercepts were included. The estimated coefficient in the model indicates that while learners were likely to select N-Num-Cl, the preference is not reliable ($\hat{\beta} = 0.99, 95\%\text{CI} = [-0.29, 2.26]$).

As in Experiment 1A, we also fitted a Bayesian hierarchical logistic regression model to test whether the two classifiers differ in the their word order patterns. The dependent variable was whether learners selected the N-Num-Cl order. The independent variables was Classifier (sum-coded). By-participant and by-item (noun in the phrase) random intercepts, by-participant random slopes for Classifier were included. Model results indicate that there is no difference in the word order between the two classifiers ($\hat{\beta} = 0.14, 95\%\text{CI} = [-1.06, 1.34]$).

We also analyzed learners’ responses in translating the artificial words into their native language. The summary of learners’ responses is shown in Table B.2 in Appendix B.2. Learners’ responses suggest that most learners did correctly acquire
Discussion

In Experiment 1B, we tested whether learners have a preference for N-Num-Cl order over N-Cl-Num order when there is no explicit evidence for the order between numeral and classifier. We found no reliable preference for either order.

However, a potential confound in this experiment is that the N-Num-Cl order at test is consistent with the order in which participants were trained on the three syntactic categories—learners were first trained on nouns, and then on numerals, and finally on classifiers. Although there was a mixed testing phase for all words before the critical testing phase in order to minimize the potential influence of block order, learners could still be biased to the N-Num-Cl order by the block order. To avoid this confound, we conduct Experiment 2, in which the block order of numeral training and classifier training were randomized for each participant.

However, in light of the lack of robust evidence for a bias for either order, in
Experiment 2, we also switch to a production task. As mentioned above, there is evidence from previous studies that biases not found in forced-choice experiments can be revealed in production tasks.

### 3.3 Experiment 2

In Experiment 2, we test whether learners have a preference between N-Num-Cl order and N-Cl-Num order when they need to produce novel phrases in the language. We used the same artificial language as in Experiment 1B, but the lexical items were presented both orthographically and orally here. In all of the testing phases, learners had to produce lexical items or phrases in the language orally.

#### 3.3.1 Methods

Experiment 2, including the hypotheses, predictions and analysis plan, was pre-registered [here](#).

**Materials**

The lexical items in Experiment 2 were the same as in Experiments 1A and 1B, but here with sound stimuli presented. Sound stimuli were recorded by a female native Mandarin speaker. The sound stimuli of all six words have the same stress pattern and syllable length. Pronunciations used for the words are indicated in Table 3.7. Image stimuli were the same as Experiment 1B.

<table>
<thead>
<tr>
<th>Syntactic category 1</th>
<th>Syntactic category 2</th>
<th>Syntactic category 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>rovo [rovo]</td>
<td>naha [naha]</td>
<td>dugu [puku]</td>
</tr>
<tr>
<td>zovo [zovo]</td>
<td>maha [naha]</td>
<td>pugu [puku]</td>
</tr>
</tbody>
</table>

Table 3.7: Experiment 2 lexicon along with pronunciations used for each word in IPA in square brackets.

111
Procedure

The procedure was similar to Experiments 1A and 1B, but was adapted for a production task.

Participants were first trained on nouns in isolation. In each trial, participants were shown an image of an object, and saw and heard the label for the object in the new language. They were instructed to repeat the word after they heard it. As in the previous two experiments, after noun training, learners were tested on what they had learned. They were shown an image of an object they had previously seen and were asked to say the label for the object. Their productions were recorded. Feedback was provided in the form of the correct response (both auditorily and orthographically) on each testing trial, regardless of whether the participant’s response was correct or incorrect. The numbers of trials in noun training and testing were the same as the previous two experiments.

The block order of numeral training and testing and classifier training and testing was randomized for each participant. After noun training and testing, learners were either trained and tested on numerals and then on classifiers (both in isolation) or on classifiers and then on numerals (both in isolation), depending on the randomized block order they received. The procedure and number of trials of numeral and classifier training and testing phases were both identical to the noun training and testing phase.

Participants were subsequently tested on all the words they had learned. The testing trial of each word was the same to that in the previous testing phase.

In the final critical testing phase, participants were instructed that they needed to use all of the three types of words they had learnt to describe the pictures. They were explicitly told that nouns in Bahareti always come first. In each trial, participants were shown an image of a certain number of objects. They were also provided with the noun, orthographically presented along with two blank spaces representing the two additional words participants needed to produce. Participants were instructed to repeat the noun first and complete the phrase in the way that they thought a native Bahareti speaker would do. Their productions were recorded. No feedback was provided.
As in the previous two experiments, after the experiment, participants were instructed to fill out a questionnaire asking them to translate the artificial words into their native language as well as their language experience.

Participants

60 native English speakers recruited through the online platform Prolific participated in Experiment 2. We used the same four pre-screening criteria as in Experiment 1. However, during data collection, we noticed that a substantial number of participants produced phrases only consisting of a numeral and a noun in the critical test. To ensure data quality, we added a pre-screening criterion that participants’ previous approval rate should be no less than 90%. We also added a reminder after the third trial of the critical testing phase to prompt participants that they need to use three words to describe the picture.\textsuperscript{14} As planned in our preregistration, participants who produced fewer than 8 analyzable data (e.g., just using two words to describe the pictures or not using the artificial language) were excluded (N=22). 7 participants with an accuracy rate below 75% in at least one of the non-critical testing phases were excluded from data analysis. 1 participant was excluded due to technical problems (i.e., no sounds in the recordings). No participants reported speaking a language with numeral classifiers. 30 participants were included in the data analysis. All participants were also paid the equivalent of £9.51 per hour for their participation.

3.3.2 Results

Vocabulary accuracy was very high (above 94%) for all three syntactic categories in Experiment 2. See Figure B.3 in Appendix B.1.2 for the results.

As in Experiment 1B, each production in the critical testing phase was coded as either only predicted by the classifier-with-numeral hypothesis or not. Figure 3.5

\textsuperscript{14}We ran a Bayesian hierarchical logistic regression model to determine whether these two changes had an influence on the results. The dependent variable was whether learners produced the N-Num-Cl order or not. The independent variable was whether the data was collected before or after having the two changes. The results of the model indicate that there is no difference in the probability of producing the N-Num-Cl order between the two subset of data.
Figure 3.5: Average proportion of phrases produced with N-Num-Cl order in Experiment 2; error bars indicate 95% confidence intervals on the mean of by-participant means; blue points indicate individual participant means; the dashed line shows chance level (50%). Learners showed no preference for N-Num-Cl order in Experiment 2.

We fitted a Bayesian hierarchical logistic regression model to determine whether the likelihood of learners producing N-Num-Cl order was reliably above chance. By-participant and by-item (image in the critical testing phase) random intercepts were included. The estimated coefficient in the model indicates that learners have no reliable preference for N-Num-Cl order ($\hat{\beta} = 1.22, 95\% CI = [-1.62, 4.33]$). Recall that in our experiment, the block order of the numeral training phase and classifier training phase is randomized for each participant. To assess whether block order had an impact on the results, we added the block order of numerals and classifiers as a predictor to the model. Model results indicate that participants who were taught classifiers first and then numerals were less likely to produce N-Num-Cl compared to those were taught numerals first and then classifiers ($\hat{\beta} = -4.08, 95\% CI = [-8.76, -0.70]$). Learners who were trained on numerals first and then on classifiers have a preference for the N-Num-Cl order ($\hat{\beta} = 6.47, 95\% CI = [1.31, 13.50]$), whereas those trained on classifiers first and then on numerals do not have such a preference ($\hat{\beta} = -1.68, 95\% CI = [-6.21, 2.24]$). This demonstrates that it is impor-
tant to consider the order in which lexical items are trained in artificial language learning experiments.

We also fitted a Bayesian hierarchical logistic regression model to investigate whether word order preferences for the two classifiers differ. The structure of the model is identical to that in Experiment 1B. Model results indicate no difference in the word order preferences for the two classifiers ($\hat{\beta} = 0.04, 95\%CI = [-1.18, 1.29]$).

As in the previous two experiments, we also examined learners’ responses in translating the artificial words into their native language in Experiment 2. Learners’ responses suggest that, most participants did correctly learn the meanings of the artificial words. The summary of learners’ responses in Experiment 2 is shown in Table B.3 in Appendix B.2.

### 3.3.3 Discussion

In Experiment 2, we tested whether learners have a preference for N-Num-Cl or N-Cl-Num orders when they needed to produce novel phrases themselves in the artificial language. We also randomised the order in which participants were trained on numerals and classifiers to ensure that any ordering preferences do not reflect training block order. Our results indicate that, again, learners showed no reliable preference for either of the two orders. Rather, learners trained with different block orders exhibited different patterns: learners who were taught numerals first and then classifiers showed a preference for N-Num-Cl order whereas those who were taught classifier first and then numerals showed no preference for either order. This demonstrates that block order does indeed impact learners’ expectation of the order of these syntactic categories. While this potentially reflects their lack of a prior preference, in Experiment 3 we train participants on numerals and classifiers within a single training block.

It is also worth noting again that a substantial number of participants in Experiment 2 produced phrases consisting of only a noun and a numeral in the critical test. This is understandable, since English does not have the grammatical category of (sortal) numeral classifiers. Moreover, numeral classifiers may appear semantically redundant, particular to English speakers, as they are indicated by the noun
itself but provide no additional information in the phrase (e.g., Craig 1992; Vittrant & Allasonnière-Tang 2021). This is particularly true in our miniature artificial language. Therefore, even though learners were instructed that the special words were needed for counting objects in Bahareti, many participants still just produced a noun and a numeral. This demonstrates another possible explanation for why participants show no clear ordering preferences: despite acquiring the lexical items, participants might still be uncertain or simply ignore classifiers as they do not add any meaning to the phrase. A potential solution to this is to test native speakers of numeral classifier languages who are familiar with this grammatical category. For example, in principle, we could test whether native speakers of a numeral classifier language in which the numeral and classifier precede the noun, have a preference for one of the two orders when both the numeral and classifier follow the noun. Unfortunately, this is not easy to do, since in most languages [Num+Cl+N] phrases exhibit word order variation. Table 3.8 summarizes variation in word order for classifier constructions in some languages for which native speakers would otherwise be relatively accessible as a participant population. In all these languages, this variation would very likely bias participants in favour of N-Num-Cl order.

However, there is another way to address the problem that sortal numeral clas-

15However, note that in numeral classifier languages, numeral classifiers can encode different information of nouns, including adding information to the nominals or supplementing information carried by nominals (Aikhenvald 2000: 319-320). Below is an example from Burmese in which numeral classifiers add information to the nominals they occur with (Becker 1975; Aikhenvald 2000: 319). The structure of the following phrases is Noun (myi, ‘river’)-Numeral (t, ‘one’)-Classifier.

(i) a. myi? to ya? ‘river one place (e.g. destination for a picnic)’
   b. myi? to tan ‘river one line (e.g. on a map)’
   c. myi? to hnuw ‘river one section (e.g., a fishing area)’
   d. myi? to sin ‘river one distant arc (e.g., a path to the sea)’
   e. myi? to khu ‘river one conceptual unit (e.g., in a discussion of rivers in general)’
   f. myi? to myi? ‘river one river (the unmarked sense)’

16”N-Num-Cl” is the order used in ancient Chinese. In Mandarin, N-Num-Cl can also be used as a nominal predication to describe the quantity of objects (Wu 2014). For example, shu san ben, bi liang zhi (lit., book three cl. pen two cl., ‘three books and two pens’).
sifiers might be neglected by native speakers of non-classifier languages. We can switch to mensural classifiers, which are found in English (e.g., slice, piece). This is what we do in Experiment 3.

### 3.4 Experiment 3

As mentioned in Section 3.1, sortal classifiers and mensural classifiers are two basic types of classifiers. Unlike sortal classifiers which may appear semantically redundant, mensural classifiers have a clear semantic contribution to the phrase by denoting the unit used to measure the quantity (e.g., Aikhenvald 2000; Vittrant & Allassonnière-Tang 2021). Below is an example from Mandarin showing the difference.

(4)  

a. *si tiao/gen huanggua*  

four cl. cucumber  

‘four cucumbers’

b. *si pian huanggua*  

four piece cucumber  

‘four pieces of cucumber’

c. *si xiang huanggua*  

four box cucumber  

‘four boxes of cucumber’

<table>
<thead>
<tr>
<th>Language</th>
<th>Dominant order</th>
<th>Order variant</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese</td>
<td>Num-Cl-N</td>
<td>N-Num-Cl</td>
<td>Aikhenvald (2000)</td>
</tr>
<tr>
<td>Korean</td>
<td>Num-Cl-N</td>
<td>N-Num-Cl</td>
<td>Aikhenvald (2000)</td>
</tr>
<tr>
<td>Mandarin</td>
<td>Num-Cl-N</td>
<td>N-Num-Cl</td>
<td>Wu (2014)</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>Num-Cl-N</td>
<td>N-Num-Cl</td>
<td>Simpson &amp; Ngo (2018)</td>
</tr>
<tr>
<td>Thai</td>
<td>N-Num-Cl</td>
<td>Num-Cl-N</td>
<td>Simpson (2008)</td>
</tr>
</tbody>
</table>

Table 3.8: Word order variation of the [Num+Cl+N] phrase in some languages.
In Mandarin, *tiao* and *gen* are sortal classifiers typically used with long and narrow things. As shown in (4a), the phrase has the same lexical meaning regardless of which sortal classifier is used. By contrast, the phrases with different mensural classifiers have different meanings: *pian* and *xiang* are distinct units for measuring quantity. Despite the differences in semantics and certain syntactic behaviours (e.g., Her & Hsieh 2010; Aikhenvald 2000; Vittrant & Allasonnière-Tang 2021), it has been argued that the two types of classifiers end up in the same syntactic position in numeral classifier languages.\footnote{Proposals vary on whether the two types of classifiers are generated in same syntactic position or not. For example, for the two types of classifiers in Mandarin, some researchers hold that they have the same syntactic structure, either combining with nouns first (Tang 2005) or combining with numerals first (Her & Hsieh 2010), whereas some other researchers consider that sortal classifiers are base-generated as the head of the ClP, and mensural classifiers are base-generated with the noun and then move upward to the head of ClP (Cheng & Sybesma 1999). However, sortal classifiers and mensural classifiers ultimately occupy the same syntactic position in these different proposals (see reviews in Her & Hsieh 2010).} This is also supported by the observation that sortal classifiers and mensural classifiers share the same word order in classifier languages (Greenberg 1972; Her 2017). Unlike sortal classifiers which only exist in some languages, mensural classifiers are much more common and exist in almost all languages. In Experiment 3, we use mensural classifiers (meaning ‘slice’) in the artificial language. As in Experiment 2, we also test whether learners have a preference for N-Num-Cl or N-Cl-Num when producing phrases in the language. The predictions remain the same.

### 3.4.1 Methods

Experiment 3, including the hypotheses, predictions and analysis plan, was pre-registered [here](#).

#### Materials

The miniature artificial language consisted of two nouns (‘pizza’, ‘cucumber’), two numerals (‘two’, ‘three’) and two mensural classifiers (one for individual complete objects and one meaning ‘slice’). In this experiment, we also adjust the lexicon slightly to reduce the similarity of the words across categories. Here we use a set
of words adapted from Martin et al. (2020). Nouns have CVC structure, and classifiers and numerals have VCVC structure. The full artificial lexicon is shown in Table 3.9. The latter two artificial word sets were randomly assigned to classifiers and numerals for each participant.

<table>
<thead>
<tr>
<th>Noun</th>
<th>Syntactic category 2</th>
<th>Syntactic category 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>uhu</td>
<td>taka</td>
<td>himi</td>
</tr>
<tr>
<td>[u’hu]</td>
<td>[ta’ka]</td>
<td>[hi’mi]</td>
</tr>
<tr>
<td>iti</td>
<td>puku</td>
<td>hono</td>
</tr>
<tr>
<td>[i’tu]</td>
<td>[bu’ku]</td>
<td>[ho’no]</td>
</tr>
</tbody>
</table>

Table 3.9: Lexicon of the miniature artificial language in Experiment 3.

The language was presented to participants both orally and orthographically. Examples of the visual stimuli are illustrated in Figure 3.6. Nouns were shown as images of either a complete object or a slice of an object (e.g., a complete pizza or a piece of pizza). To train participants on classifiers, participants were always simultaneously shown two images of the same noun meaning (e.g., pizza), one of which was highlighted while the other was greyed out. Objects in the two images differed only in measure unit, not number (e.g., two pieces of pizza vs. two whole pizzas). The particular classifiers meaning being trained was illustrated by the measure unit of objects in the highlighted image. Numeral meanings were also taught to participants using this same trial structure. Two images with same noun meaning (e.g., cucumber) were shown, one of which was highlighted while the other was greyed out. Objects in the two images always differ in number only, not measure unit (e.g., two whole cucumbers and three whole cucumbers). The particular numeral meaning being trained was illustrated by the number of objects in the highlighted image.

To encourage participants to pay attention to the instructions, at the beginning of each phase the instructions were presented both auditorily and orthographically. The auditory instructions were recorded by a female native English speaker.

**Procedure**

Participants were first trained and tested on nouns in isolation (8 trials for each of the two nouns, presented in random order with no more than two consecutive trials for the same noun). They were instructed that they would learn the labels
for some objects in Bahareti. For each noun, participants were shown an image of a complete object in half of the trials and an image of a slice of that object in the other half. The procedure was the same as Experiment 2 except that (i) in the training trials, before the appearance of the artificial word and the image, the phrase “object label” appeared first at the top left corner of the screen to show the type of the artificial word that would appear in this trial; (ii) in the testing trials, the phrase “object label” also appeared first at the top left corner and then the image appeared.

Participants were then trained and tested on numerals and classifiers in isolation (8 trials for each of the four words, presented in random order, with no more than two consecutive trials for the same word). They were instructed that in the following trials they would learn two types of words that are needed for describing the quantity of objects in Bahareti. For the numeral training, learners were first shown the phrase “number word” at the top left corner of the screen to indicate the type of word they were learning. They then saw and heard the number word simultaneously. They were then shown two images as described above: the unit to measure the quantity of the objects was the same but the number of the objects was different (e.g., one image of two whole cucumbers and one image of three whole cucumbers). First, both images were greyed out. Then the left image was un-greyed out and highlighted and the word for the number of objects in the left image appeared. Participants were instructed that the word expresses the number of objects in the highlighted image. Participants were instructed to try to learn the word and
then press the spacebar to continue. Then, the left image was greyed-out and the right image was un-greyed out and highlighted. The word for the number of objects in the right image appeared. Positions of the two images were randomized for each trial. The training of mensural classifiers was exactly the same, except that participants were shown two images in which the number of objects was the same but the unit to measure the quantity of the objects was different (e.g., one image of two slices of pizza and one image of two whole pizzas). For these trials, the phrase "measure word" was shown at the top left corner of the screen. Participants were instructed that the word expresses the measure unit of objects in the highlighted image. The order of numerals and classifier training trials was randomized for each participant.

After the training trials, there was a matching test phase for numerals and classifiers (4 trials for each of the four words, presented in random order, with no more than two consecutive trials for the same word). Participants were shown two images in each trial. The number or measure unit (but not both) of the objects in the two images differed. Participants heard and saw a word describing the number or measure unit of objects in one of the two images. Participants were asked to click on the image that matched the word. Feedback was provided on each matching trial. The correct image matching the word was marked with a tick whereas the image not matching was greyed-out and marked with a cross. At the same time, there was a correct sound if the participant’s answer was correct, and an incorrect sound if the participant’s answer was incorrect.

After the matching test, participants were further tested on the production of numerals and classifiers (4 trials for each of the four words, presented in random order, with no more than two consecutive trials for the same word). Participants were first shown the phrase "number word" (for numerals) or "measure unit" (for mensural classifiers) at the top left corner of the screen. They were then shown two greyed-out images that they had previously seen in numeral or classifier training trials. Then the left image was un-greyed out and highlighted and they were instructed to say the number or the measure unit of the objects in that image. Their production was recorded. After saying the word, participants were instructed to press the spacebar to continue. Feedback was provided in the form of the correct
response (both auditorily and orthographically) on each testing trial, regardless of whether the participant’s response was correct or incorrect. Then, the left image greyed out again and the right image was ungreyed-out and highlighted. Participants were instructed to say the number or measure unit of the objects in the right image. Positions of the two images were also randomized for each trial.

Participants were subsequently tested on all of the six words they had learnt, using the same procedure.

Participants were lastly tested on the phrases consisting of the three syntactic categories (2 trials for each of the eight images). The procedure was the same as Experiment 2. Participants were instructed that they needed to say the object label first and then complete the phrase with the other two words.

As in the previous three experiments, after the experiment, participants were prompted to complete a questionnaire asking them to translate the artificial words into English and provide information about their language experience.

Participants

60 native English speakers recruited through the online platform Prolific participated in Experiment 3. We used the same four pre-screening criteria as in Experiment 2. As planned in our preregistration, participants who produced less than 12 analyzable trials were excluded (N=5). 24 participants with an accuracy rate below 75% for at least one of the three syntactic categories in the mixed testing phase were excluded from data analysis. 1 participant was excluded due to technical problems (i.e., no sounds in the recordings). No participants reported speaking a language with numeral classifiers. 30 participants were included in the data analysis. All participants were paid the equivalent of £10.41 per hour for their participation.

3.4.2 Results

Vocabulary accuracy for each syntactic category in the mixed testing phase in Experiment 3 was very high (above 95%). See Figure B.4 in Appendix B.1.3 for the full results.
Each production in the critical testing phase was coded as either only predicted by the classifier-with-numeral hypothesis or not. Figure 3.7 shows the proportion of producing the N-Num-Cl order in Experiment 3.

We fitted a Bayesian hierarchical logistic regression model to determine whether the likelihood of learners producing N-Num-Cl order is reliably above chance. By-participant and by-item (image in the critical testing phase) random intercepts were included. The estimated coefficient in the model indicates, once again, that while learners were likely to produce N-Num-Cl order, the preference is not reliable ($\hat{\beta} = 1.19, 95\%CI = [-0.95, 3.47]$).

As in previous three experiments, we fitted a Bayesian hierarchical logistic regression model to investigate whether the word order preference for the two classifiers differ. The structure of the model is the same as for Experiments 1B and 2. Model results indicate that there is no difference between the two classifiers ($\hat{\beta} = -0.11, 95\%CI = [-0.71, 0.48]$).

In light of the fact that in both Experiments 2 and 3 so far there is a slight but non-reliable preference in favour of N-Num-Cl, we also ran an analysis of the combined
data from both experiments. We fitted a Bayesian hierarchical logistic regression model predicting whether learners produced N-Num-Cl order as the dependent variable. The independent variable was Experiment (sum-coded). By-participant and by-item (image in the critical testing phase) random intercepts were included. The conclusion of this model remains the same: learners were likely to produce N-Num-Cl order, but this preference is not reliable ($\hat{\beta} = 1.49$, 95%CI = [-1.33, 4.53]).

3.4.3 Discussion

In Experiment 3, we tested whether learners have a preference for N-Num-Cl or N-Cl-Num orders with mensural classifiers. Our results indicate that again, learners showed no reliable preference between the two orders.

As in the previous three experiments, we also examined learners’ responses in the translation task in the post-experimental questionnaire. Learners’ responses suggest that almost all of the participants did correctly learn the meanings of the words. The summary of learners’ responses in Experiment 3 is shown in Table B.4 in Appendix B.2. As mentioned earlier, mensural classifiers used in Experiment 3 are common across languages. If learners spoke a second language in which the numeral follows the noun (e.g., Lao and Yoruba), it is possible that learners’ production is influenced by the order used in their second language. To exclude this possibility, we also examined participants’ second language experience. Learners’ responses show that no participants spoke a second language with N-Num order.

Interpreting null results is challenging; nonetheless, the absence of a consistent preference across participants in Experiment 3 might indicate that there is no universal or preferred underlying structure for mensural classifiers. As with sor-tal numeral classifiers, two different underlying structures have been proposed for phrases with mensural classifiers. For example, Rothstein (2009, 2011) argues that mensural classifiers in English such as piece, and item first combine with nouns to yield a set of parts and then combine with numerals (i.e., the partition structure). The mensural classifier is the head of this phrase. By contrast, mensural classifiers such as pound and gram specify the unit of measurement and combine with numerals first (i.e., the measurement structure). In this structure, the noun denoting the
object being measured is the head. This analysis has been supported by the varied syntactic behaviors of two types of mensural classifiers (Bale et al. 2019). For example, the count/mass status of the phrase with different mensural classifiers is different:

(5)  a. Many of / #much of the 300 items of furniture will be exported to India.  
    (Bale et al. 2019: 7)

   b. Much of / #many of the 300 tonnes of furniture will be exported to India.  
    (Bale et al. 2019: 7)

It is unnatural to use *much* to modify the structure with a partition reading in which the head is the count mensural classifiers, whereas it is unnatural to use *many* to modify the phrase with a measurement reading in which the head is the mass noun, as shown in (5). Likewise, the two types of mensural classifiers also differ in number agreement:

(6)  a. The 500 items of furniture that we bought at the market were / #was in good condition.  
    (Bale et al. 2019: 7)

   b. The 500 grams of sugar we added to the sauce gives / ?give it a sweet aftertaste.  
    (Bale et al. 2019: 7)

It is more natural to use plural agreement on the verb when the subject is a partition structure, but singular agreement when the subject is a measurement structure, as shown in (6).

Moreover, even for the mensural classifiers such as *piece* and *slice*, evidence from constituency tests, which is a standard method to identify the hierarchical structure, is also mixed (for example a lengthy back-and-forth discussion on numeral classifiers can be found on the LINGUIST list running from 1993 to 1994). In some contexts, the results suggest that mensural classifiers can behave as a syntactic unit with both nouns and numerals. For example, a coordination sentence is grammatical when *either* the combination of mensural classifiers and nouns is deleted or the noun is deleted, as shown in (7)\footnote{The examples in (7), (8), and (9) are based on examples from Edith Moravcsik in the discussion on classifiers on the e-mail LINGUIST list from 1993 to 1994.}:

125
I ate three pieces of pizza, and he ate four (pieces).

Likewise, it is also natural to use numerals or the combination of numerals and mensural classifiers to answer questions, as shown in (8).


However, in some other contexts, constituency tests suggest that numerals and mensural classifiers behave as a syntactic unit. For example:

(9) a. ?two pieces and three pieces of pizza makes five pieces of pizza
   b. *two and three pieces of pizza makes five pieces of pizza

(10) I ate two slices of ham, and several of bread. 19

The deletion test in (9) suggest that in a coordination structure, the sentence remains acceptable after removing nouns; however, the sentence is ungrammatical after removing both mensural classifiers and nouns. This suggests that the numeral and mensural classifier likely form a constituent here. Similarly, the substitution test in (10) suggests that it is grammatical to use a word as a substitute for the combination of numeral and mensural classifier; however, the combination of mensural classifier and noun cannot be substituted by a pro-form in this context.

In summary, learners might have experience with both possible structural representations of mensural classifiers. They might also have different individual experiences with the usage of classifiers in diverging contexts. In a situation where there is no evidence for the underlying structure of the phrase in a new language, learners’ individual experience of the use of mensural classifiers in their native language could lead them to choose one of the two possible structures. Interestingly, in our results, while substantial variation between participants is observed, within each participant, the variation is limited. This is illustrated by the bimodal distribution of data points in Figure 3.7. Specifically, more than half of the learners in our experiment preferred the order in which classifiers and numerals are in the same

19This example is based on an example from Paul Kershaw in the discussion on classifiers on the e-mail LINGUIST list from 1993 to 1994.
syntactic unit (i.e., N-Num-Cl), and roughly one-third of the learners preferred the order which is possible under both analyses (i.e., N-Cl-Num). By contrast, only two participants showed strong variability in their data, with the frequency of the two orders in their productions being similar. This bimodal distribution of the results aligns with the discussion above: learners might, based on their individual experiences, choose one of the two orders which are either possible or not under the two different underlying structures in the experiment.

3.5 General discussion

3.5.1 Numeral classifiers: cross-linguistically heterogeneous

Numeral classifiers are an important typological feature of many languages. Their semantic and syntactic status, however, has long been disputed. One view holds that numeral classifiers are required for the interpretation of numerals in classifier languages, and form a syntactic constituent with them. By contrast, an alternative view holds that they are required for the interpretation of nouns and form a syntactic constituent with them. While there are a number of sources of evidence for each theory, word order of the three syntactic categories has been taken as a crucial piece of evidence to adjudicate between them: only word order patterns that are predicted by the classifier-with-numeral hypothesis (i.e., orders in which numerals and classifiers are adjacent) are observed across classifier languages. However, the geographic distribution of numeral classifiers is highly constrained and the typological tendency of classifier order likely reflects the result of language contact. Converging evidence indicating a preference for the word order patterns that are predicted by one of the analyses is therefore needed. Here, we use artificial language learning experiments to bring new behavioural evidence to bear on this question. Specifically, in Experiment 1A, we tested whether learners have a preference between (i) Cl-N-Num order vs. N-Num-Cl order, or between (ii) Num-N-Cl order vs. Cl-Num-N order. In Experiments 1B and 2, we tested whether learners have a preference between (iii) N-Num-Cl and N-Cl-Num order. In Experiment 3, we tested whether learners have a preference between these same orders when the clas-
sifiers are mensural rather than sortal. While in almost all experiments there was if anything a slight trend favouring orders predicted by the classifier-with-numeral hypothesis, this preference was not reliable.

While the results of this set of experiments are therefore inconclusive, the consistent lack of bias across all of the experiments points to the possibility that there might be no single preferred underlying structure for the [Num+Cl+N] phrase. As discussed in Section 3.1.1, a number of researchers indeed hold that the syntactic status of numeral classifiers is variable, and differs across languages (e.g., Jenks 2010; Gil 2013; Bale et al. 2019; Little et al. 2022). According to this view, classifiers in some languages form a constituent with numerals. In such languages, the occurrence of classifiers depends on numerals. For example, in Ch’ol (a Mayan language), there are two types of numerals: Mayan-based numerals and numerals borrowed from Spanish. These two types of numerals behave differently in terms of whether a numeral classifier is required when numerals combine with nouns: Mayan-based numerals require a classifier, whereas numerals borrowed from Spanish do not (Little et al. 2022). Below are two examples showing this difference (Bale & Coon 2014: 701):

(11) a. *ux-*(p’ej) tyumuty
    three-cl. egg
    ‘three eggs’

b. nuebe-(*p’ej) tyumuty
    nine-cl. egg
    ‘nine eggs’

The Mayan-based numeral ux requires the classifier, whereas the numeral nuebe, which is borrowed from Spanish, resists the classifier, as shown in (11). Moreover, in Ch’ol, all classifiers are only used with numerals, and classifiers are always required by Mayan-based numerals, even when counting and referring to the number

---

20In Experiment 1A, we observed a weak preference for Num-N-Cl in a narrower confidence interval in Condition 2. However, as discussed in Section 3.2.1, the result could be confounded by the fact that some participants made their choice based on the position of numeral. We therefore interpret this as not very robust.
The syntactic behaviors of the two types of numerals suggest that classifiers are required by numerals in Ch’ol, and form a syntactic unit with classifiers in this language (Bale & Coon 2014; Little et al. 2022).

However, in some other languages, the occurrence of numeral classifiers depends on nouns. For example, in Vietnamese, some nouns optionally take classifiers, some nouns required classifiers, some nouns, however, resist classifiers (Simpson & Ngo 2018; Dékány 2021). Below are some examples (Simpson & Ngo 2018: 213-214):

(12) a. hai *(con) chó
two cl dog
‘two dogs’
b. tám (cài) làng
eight cl village
‘eight villages’
c. hai màu
two color
‘two colors’

The data from these different languages suggest that numeral classifiers may be syntactically and semantically heterogeneous across languages. In other words, there may be no universal basis for the structure of classifiers. If this is the case, then we would expect that without sufficient evidence for the structure of classifiers in a new language, learners may arbitrarily choose an order, or a structure. This is markedly different from previous work testing ordering preference for complex noun phrases consisting of a demonstrative, a numeral, an adjective, and a noun. In this case, there is a single underlying structure that has been argued for in the theoretical literature (typically argued to be derived from universal semantic scope relations). Orders that reflect this hierarchical structure are more common typologically, and preferred by learners in experimental tasks similar to the ones used here (Martin et al. 2020, 2024). In this case there is converging evidence for a
universal hierarchical structure; in the case of classifiers it may simply be the case that no such universal structure exists.

It’s also worth mentioning that the results of our experiments are strikingly different from what has been reported for experiments testing preferences for the order of noun stems with number and gender morphemes. Recall that for the latter, Saldana & Culbertson (2023) observed a strong preference to place gender closer to the noun base than number. While the two cases differ in that one targets word order and the other morpheme order, the difference in the results is still somewhat unexpected. Both concern the order of elements expressing noun, numerosity, and nominal category (at least in the case of sortal classifiers). In our experiments a given noun can only occur with a sortal classifier; in Saldana & Culbertson (2023), however, a noun base occurs with both gender morphemes. From this point of view, the semantic association between noun and classifier in our experiments is stronger than that between gender and noun base in Saldana & Culbertson (2023). Similarly, in the experiments testing the order of nouns with numerals and adjectives mentioned above (e.g., Culbertson & Adger 2014; Culbertson et al. 2020b; Martin et al. 2020), a noun occurs with different adjectives. Given that a bias for ordering adjective/gender closer to noun (base) than numeral/number is observed in the two studies, we would therefore have expected similar results (i.e., a bias for ordering classifiers closer to nouns than numerals) in our experiments. One factor which may explain the difference between the results of our experiments and previous studies is that gender morphemes in Saldana & Culbertson (2023) and adjectives in Martin et al. (2020) have semantic contribution to the whole word or phrase. Specifically, gender morphemes in at least a subset of the experiments reported in Saldana & Culbertson (2023) lead to a difference in the meaning of the nouns (i.e., the references of nouns differ based on the use of different gender morphemes) and adjectives in the experiments reported in Martin et al. (2020) also lead to different meanings of the phrase (i.e., the references of phrases differ based on the use of different adjectives). By contrast, in our experiments, the sortal numeral classifiers do not have this kind of semantic contribution to the phrase. This varying degree of semantic contribution of the gender morphemes and adjectives in previous work and the numeral classifiers in current experiments might lead to the different results.
3.5.2 Disparity in typology and language learning

Regardless of the exact explanation for why the result of our experiments differ from previous experiments, they suggest that there may not be any direct causal link between cognitive biases and typological patterns of word order in the [Num+Cl+N] phrase. Table 3.10 summarizes the typological frequency of orders tested in each experiment and results of each experiment.

<table>
<thead>
<tr>
<th>Experiment tested</th>
<th>Word order tested</th>
<th>Typologically more frequent order</th>
<th>Order preferred by learners</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp. 1A</td>
<td>Cl-N-Num, N-Num-Cl, Num-N-Cl, Cl-Num-N</td>
<td>N-Num-Cl</td>
<td>No reliable preference</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cl-Num-N</td>
<td>A weak preference for Num-N-Cl</td>
</tr>
<tr>
<td>Exp. 1B</td>
<td>N-Num-Cl, N-Cl-Num</td>
<td>N-Num-Cl</td>
<td>No reliable preference</td>
</tr>
<tr>
<td>Exp. 2</td>
<td>N-Num-Cl, N-Cl-Num</td>
<td>N-Num-Cl</td>
<td>No reliable preference</td>
</tr>
<tr>
<td>Exp. 3</td>
<td>N-Num-Cl, N-Cl-Num</td>
<td>N-Num-Cl</td>
<td>No reliable preference</td>
</tr>
</tbody>
</table>

Table 3.10: Summary of the typological frequency of orders tested and results in each experiment.

As shown in Table 3.10, in no case did learners show any reliable preference for the order that is relatively more frequent in typology. This huge disparity in typology and language learning, which stands in contrast to many other artificial language learning experiments, suggests that the distribution of numeral classifiers at the global level may instead reflect cognition-external factors, in particular language contact. In fact, compared to other noun classification systems such as gender, classifiers are more easily diffused across different language families and less stable in terms of historical inheritance within language families (e.g., Nichols 1992; Seifart 2010; Greenhill et al. 2017; Allassonnière-Tang & Dunn 2020; Allassonnière-Tang et al. 2021). For example, numeral classifiers are remarkably similar in East

21This is also consistent with the prediction that content words are more likely to be borrowed than functional words in language contact (see Seifart 2010).
and Southeast Asia (Bisang 1999) and has been taken as a feature of the Mainland Southeast Language Area (e.g., Enfield 2005; Comrie 2007). Indeed, a number of studies suggest that numeral classifiers in certain regions are products of language contact. For example, Nichols (1992: 251) points out that the distribution of numeral classifiers "... only one clear hotbed and it lies along the circum-Pacific migration route, strongly suggests that entrants to the New World came from a single population characterized by high frequency of numeral classifiers (except, apparently, for the earliest entrants, whose descendants are now mostly in southern South America and eastern North America, where numeral classifiers are rare)". Based on a detailed comparison of the function and word order of numeral classifiers across different language groups, Her & Li (2023) argue that numeral classifiers in Asia and the Pacific just developed initially in one language group (likely Sinitic) and later spread to all other classifier languages through language contact. By using the Bayesian phylogenetic inference method, Allasonnière-Tang et al. (2023) argue that the numeral classifiers in the Turkic, Mongolic, and Tungusic languages of the Altaic region are also likely borrowed from other languages (Allasonnière-Tang et al. 2023). It has also been argued that numeral classifier in Mayan languages are also a consequence of language contact (Law 2020). If this is the case, then the distribution of classifier order across languages reflects the relatedness of classifier languages, but not a valid statistical universal.

---

22 It has also been pointed out by Janhunen (2000: 705) that "a feature like the numeral classifiers in East Asia can hardly have originated separately in several adjacent languages and language families. Rather, there was a single primary innovation where the principle was first created, and from where it radiated to all over the region."
Chapter 4

Summary and conclusion

4.1 Summary and contributions

Since the seminal work by Greenberg (1963), many statistical universals have been generalized based on large-scale cross-linguistic comparisons. In the past few decades, different hypotheses have been proposed to account for the origin of typological universals. For example, typological universals may be reflections of innate Universal Grammar, or results of cognitive biases in language learning and use, or results of diachronic changes. In some cases, cross-linguistic tendencies may reflect the genealogical or geographic relatedness of languages in the samples. These explanations have been the subject of much debate. One possible reason is that evidence for these different explanations mainly comes from typological and historical data. However, establishing causal relations between specific hypothesized mechanisms and particular language features is challenging when relying solely on these data. In this thesis, I focus on the role of language learning in shaping language universals. I take an artificial language learning approach to study the cognitive basis of two word order universals: cross-category harmony (Part I, Chapter 2) and the order of numeral classifiers (Part II, Chapter 3). By teaching learners miniature artificial languages that are precisely designed for the relevant language features, causal relations between language learning and the two word order universals were directly tested.
4.1.1 Summary of experiments

In Chapter 2, I investigated the cognitive basis of cross-category harmony, one of the most well-known language universals in linguistics. Cross-category harmony describes the tendency for word order to correlate across different types of phrases within languages. However, it has long been disputed why harmony holds and what it holds between. Accounts raised in the literature include hard constraints on syntactic representations (e.g., the Head-directionality parameter as in Travis 1984), a domain-general bias for simplicity (e.g., Culbertson & Kirby 2016), a processing bias for dependency-length minimization (e.g., Hawkins 1994), a shared historical origin (e.g., Collins 2019), or accidents of history (e.g., Dunn et al. 2011).

I conducted three experiments to investigate two questions. First, do learners have a preference for particular word order patterns, and do learners’ preferences align with the tendencies for harmony observed in typology? Specifically, I tested word order harmony between two pairs of phrases which play an important role in typology: harmony between verb phrase and adpositional phrase (Experiment 1), which has strong typological evidence, and harmony between verb phrase and noun phrase consisting of a (color) adjective and a noun (Experiment 2), which has been rejected by large-scale typological surveys with different methods. I found a similar pattern in language learning: learners had a strong preference for harmony between verb phrase and adpositional phrase regardless of whether they spoke a language with harmonic orders (English speakers) or mixed orders (Mandarin speakers) in the two domains, but no preference for harmony between verb phrase and noun phrase consisting of a color adjective and a noun. I then asked the second question: why does the bias for harmony differ across these phrase types? The hypothesis is that the presence/absence of a harmony bias between different phrase types is sensitive to the similarity of the two syntactic categories to be aligned in the phrases. I then conducted Experiment 3 to compare learners’ preference for aligning verbs (in verb phrases) with adjectives (in noun phrases) when adjectives are semantically more and less similar to verbs (i.e., active adjectives (e.g., ‘broken’) vs. stative adjectives (e.g., ‘red’, ‘stripy’)). Results confirmed the hypothesis, showing that when adjectives are active and more verb-like, learners exhibited a bias to
align verbs with adjectives; when adjectives are stative and less verb-like, however, learners did not show such a bias.

The results of this series of experiments reported in Chapter 2 have three key conclusions. First, the learning bias for harmony between the two pairs of phrases roughly matches the tendency in typology (either harmony or lack-ther eof). Since no grammaticalization pathways and no combinations of phrases were involved in the experiments, learners’ bias to align verbs (in verb phrases) with adpositions (in adpositional phrases) observed in Chapter 2 cannot be explained by alternative explanations for harmony in the literature, such as the diachronic accounts and the processing bias. Instead the results suggest that harmony reflects a cognitive bias, active during language learning, for aligning similar elements across phrases. Previous research using artificial language learning has suggested this for different nominal modifiers (Culbertson et al. 2020a). However, crucially, these experiments provide the first evidence to tease apart different explanations for cross-category harmony, suggesting that this classic word order universal is at least partly shaped by human cognition. The results of this series of experiments contribute evidence to the broader debate on whether language universals are generally driven by cognitive factors (e.g., Christiansen & Chater 2008; Kiparsky 2008) or solely influenced by language change (e.g., Blevins 2004; Collins 2019).

In addition to this evidence for the connection between cognition and cross-category harmony, these results also have implications for theories of the specific mechanism underlying harmony. In Experiment 3, learners showed a preference to align verbs with active adjectives but not with stative adjectives. This demonstrates that a bias for harmony is likely not purely based on syntactic concepts such as head and dependent—the most influential explanation in the literature—but is also influenced by the semantic similarity of both categories to be aligned. While it has been discussed in previous studies that semantic similarity might play a role in shaping harmony, the results of Experiment 3 provide the first empirical evidence for this. This idea links the cognitive mechanisms responsible for harmony to those responsible for recurrent pathways of diachronic change. Specifically, both harmony and recurrent grammaticalization pathways might be influenced by the similarity of different categories. The probability of grammaticalization between
two categories (e.g., verbs to adpositions) and the strength of the harmony bias between them (e.g., aligning verbs with adpositions) may both reflect the degree of similarity between them.

Finally, in the experiments reported in Chapter 2, while non-harmonic orders were generally dispreferred by learners, some participants did learn them. This underscores the fact that non-harmonic orders are not impossible—already suggested by the typological evidence. This supports the claim that harmony is likely the result of a soft bias rather than a hard constraint. Such hard constraints are regularly proposed for related word order patterns in the theoretical linguistics literature. For example, the Final-over-Final Constraint was first proposed by Biberauer et al. (2014), as a ban on orders in which a head-initial phrase is immediately dominated by a head-final phrase. This would rule out as impossible to learn the combination of OV with PrepN order. Yet, in the OV condition of both Experiments 1 and 3, a few learners appear to have learned exactly this order. The idea that cross-category harmony is instead a soft bias is in line with the growing body of literature on the role of a general bias for simplicity in shaping language structure (e.g., Culbertson et al. 2012; Smith et al. 2013; Durvasula & Liter 2020; Carr et al. 2020), providing evidence for the view that language universals are likely the results of interaction between domain-general cognitive mechanisms and language representations (e.g., Culbertson & Kirby 2016; Ferdinand et al. 2019).

In Chapter 3, I investigated the cross-linguistic tendencies in the order of numeral classifiers. Numeral classifiers are one of the most frequent nominal classification systems across language. It has been observed that for the phrase consisting of a numeral, a classifier, and a noun, only orders in which numeral classifiers are adjacent to numerals are attested (e.g., N-Num-Cl). This has been taken as one of the most important pieces of evidence distinguishing between alternative analyses of the syntactic status of numeral classifiers: whether the hierarchical structure underlying these phrases is such that numeral classifiers combine with numerals first, or combine with nouns first. However, typological evidence is not very reliable here since numeral classifiers are tightly constrained to specific geographic regions and likely reflect language contact. I conducted three artificial language learning experiments to search for new sources of evidence for this issue. I investi-
gated two questions: (i) do learners have a bias favoring word order patterns that are predicted by one of the two competing analyses mentioned above? and (ii) do learners’ biases match the tendency we observe in typology? Specifically, I tested whether learners have a preference for (i) Cl-N-Num order vs. N-Num-Cl order, or (ii) Num-N-Cl order vs. Cl-Num-N order (Experiment 1A), or for (iii) N-Num-Cl vs. N-Cl-Num order (Experiment 1B with a forced-choice task and Experiment 2 with a production task). I also tested whether learners have a preference for one of these same orders (i.e., N-Num-Cl vs. N-Cl-Num) when the classifiers are mensural rather than sortal (Experiment 3). The two competing hypotheses make different predictions about which order should be preferred by learners in these experiments.

In almost all experiments, I found, if anything, a subtle trend favouring orders predicted by the classifier-with-numeral hypothesis. However, this preference was not reliable, either within individual experiments, or collapsing across different experiments. This result potentially aligns with previous literature suggesting that the syntactic status of classifiers is language-specific: the syntactic and morphological behaviors of classifiers vary across languages. For example, in some languages classifiers are required for some numerals but not for others, whereas in some other languages classifiers are required for some nouns but not for others. I argued that the consistent absence of a reliable preference across the set of experiments in Chapter 3 might suggest that indeed there is no universal or preferred underlying structure of the [Num+Cl+N] phrase. In other words, numeral classifiers might be syntactically heterogeneous across languages and both the structure predicted by the classifier-with-numeral analysis and that predicted by the classifier-with-noun analysis are possible. The huge disparity in typology—where just a few orders are very common—and learning—where there was no reliable preference—implies that the cross-linguistic tendency in the order of numeral classifiers may be driven by factors beyond cognition. In this case, language contact appears to be a likely candidate explanation.

A summary of all the experiments reported in this thesis is presented in Table 4.1.
<table>
<thead>
<tr>
<th>Word order</th>
<th>Experiment</th>
<th>Word order tested</th>
<th>Typological tendency</th>
<th>Learning bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-category harmony</td>
<td>Exp. 1</td>
<td>Harmony between VP and PP</td>
<td>A clear tendency for harmony</td>
<td>A clear learning bias for harmony</td>
</tr>
<tr>
<td>Exp. 2 and 3</td>
<td></td>
<td>Harmony between VP and NP</td>
<td>No tendency for harmony</td>
<td>No harmony bias when adjective are stative A harmony bias when adjectives are active</td>
</tr>
<tr>
<td>Numeral classifier order</td>
<td>Exp. 1A</td>
<td>Cl-N-Num, N-Num-Cl</td>
<td>N-Num-Cl is much more common</td>
<td>No reliable preference</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Num-N-Cl, Cl-Num-N</td>
<td>Cl-Num-N is more common</td>
<td>A weak preference for Num-N-Cl</td>
</tr>
<tr>
<td></td>
<td>Exp. 1B</td>
<td>N-Num-Cl, N-Cl-Num</td>
<td>N-Num-Cl is much more common</td>
<td>No reliable preference</td>
</tr>
<tr>
<td></td>
<td>Exp. 2</td>
<td>N-Num-Cl, N-Cl-Num</td>
<td>N-Num-Cl is much more common</td>
<td>No reliable preference</td>
</tr>
<tr>
<td></td>
<td>Exp. 3</td>
<td>N-Num-Cl, N-Cl-Num</td>
<td>N-Num-Cl is much more common</td>
<td>No reliable preference</td>
</tr>
</tbody>
</table>

Table 4.1: A summary of all the experiments reported in this thesis.

### 4.1.2 General conclusions

In general, the work presented in this thesis shows that cognition plays different roles in shaping cross-linguistic tendencies. Some cross-linguistic tendencies, such as cross-category harmony, are at least partly shaped by features of human cognition; some other cross-linguistic tendencies, such as the word order of numeral classifiers, are more likely to be results of cognitive-external factors, such as language contact. Interestingly, the (in)consistency between the patterns observed in
learning and typology regarding the different language features may correspond to the reliability of the typological tendencies themselves. For cross-category harmony, which is one of the most robust universals, with evidence from large-scale typological surveys with different methods (e.g., Dryer 1992; Jäger & Wahle 2021), we observed a consistent cognitive universal; for the typological tendency of classifier order, which is likely not very reliable given that the language sample is not very large (Greenberg 1972) and the relatedness of languages in the sample has not been effectively controlled, we did not observe a consistent cognitive universal. These results demonstrate that the artificial language learning approach is an effective tool to test hypotheses about the potential cognitive underpinnings of typological universals. Importantly, this includes adjudicating between very distinct candidate explanations, like diachrony vs. learning, and explanations that differ only in terms of the details of the analysis proposed, i.e., whether classifiers form a constituent with nouns or numerals. In addition to providing evidence for the origin of language universals, behavioral data from artificial language learning experiments is thus an important supplement to typological data. Converging evidence from different sources can provide more reliable support for (or against) linguistic universals.

4.2 Future directions

In the section above, I summarized the conclusions and contributions of this thesis. However, there is still much work to be done in these areas. Below I will provide a brief discussions of what I see as promising future directions for each.

4.2.1 Measuring between-category similarity

In Chapter 2, I manipulated the semantic similarity between verbs and adjectives to test the hypothesis that the cognitive bias for harmony is sensitive to semantic similarity of different syntactic categories. The results of Experiment 3 confirmed this hypothesis. I discussed that the strong bias for alignment between (active) verbs and adpositions, as opposed to no bias for alignment between (active) verbs
and stative adjectives, also likely reflects the similarity the holds between adpositions and (active) verbs, in contrast with stative adjectives. In Chapter 2, I provided some syntactic and semantic justifications for this view. However, some methods in natural language processing, such as vector-based similarity measures, allow us to evaluate the similarity of different syntactic categories quantitatively. In addition to making predictions about harmony, this kind of data could also make predictions about whether a syntactic category is more likely to grammaticalize into another category. In the future, it would be highly beneficial to use these methods to test whether verbs are indeed more similar to adpositions than adjectives. The findings of such an investigation could offer converging evidence for the view that the harmony bias is influenced by cross-categorical similarity, as well as help test the idea that the underlying mechanism for harmony and diachronic change is the same.

4.2.2 Extending to populations with fully non-harmonic native languages

In Chapter 2, to assess harmony between the verb phrase and adpositional phrase, I conducted tests on two populations: native English speakers and Mandarin speakers. English has harmonic orders in the two domains (i.e., VO and PrepN), and Mandarin has mixed orders (i.e., VO, both PrepN and NPost). As reported in Chapter 2, there is a clear harmony bias in both populations and no difference in the strength of the harmony bias was observed between the two populations. Since Mandarin speakers who have experience with non-harmonic orders also have a clear bias for harmony, these results are hard to explain solely by L1 influence; instead, they indicate that the harmony bias is likely a universal cognitive bias. However, Mandarin also has harmonic orders in the two domains. The bias observed in Mandarin speakers could still be influenced (to some extent) by learners’ experience of harmonic orders in their native language. The strongest evidence for a universal bias for harmony therefore should come from native speakers of languages with fully non-harmonic orders in the two domains, such as Persian. Persian is an OV language with prepositions (Dryer 2013h). In the future, it is worth testing native speakers of such languages. If a harmony bias is observed in these populations,
it will provide the strongest evidence for a universal cognitive bias for harmony.

4.2.3 Harmony based on alignment of lexical/phrasal categories

Another question for future research is testing the explanatory role of the Branching direction theory (Dryer 1992). As introduced in Chapter 2, while the syntactic notions of head and dependent are the most influential theory regarding what harmony holds between (i.e., heads align with heads), there is also an alternative explanation, the Branching direction theory, which argues that harmony is not based on the alignment of heads, but instead on alignment of categories that share the property of being phrasal or not (i.e., consisting of a single lexical category or a phrase). According to this theory, lexical categories align with the verb in the verb phrase (which is also a lexical category) and phrasal categories align with the object in the verb phrase (which is also a phrasal category). In Experiment 3 in Chapter 2, learners showed a tendency to align verbs with adjectives only when the adjectives are active and thus more similar to verbs. Given that both stative and active adjectives in the experiment are lexical categories, these results suggest that a harmony bias is not purely based on alignment of lexical/phrasal categories. However, it remains unclear how the phrasal or lexical properties of elements to be aligned would influence the presence or absence of a harmony bias. For example, it would be of interest to test whether, all things being equal, learners tend to align lexical categories but not phrasal categories with verbs, and align phrasal categories but not lexical categories with objects. This would suggest that both semantic and syntactic or distributional similarity play a role in driving harmony.

4.2.4 The learning bias and processing bias: interaction and competition

A different explanation for cross-category harmony that I reviewed in Chapter 2 is a preference for dependency-length minimization. This theory argues that word orders in which the dependency distance between (heads and) dependents is minimized are more efficient to parse (Hawkins 1994, 2004). Since this processing ex-
planation makes predictions about combinations of phrases, and learners were only exposed to phrases consisting of two words in the experiments in Chapter 2, the harmony bias observed in the experiments in Chapter 2 is not likely to be the result of this processing bias. However, it would be intriguing to examine how learners behave when exposed to combinations of phrases where a processing bias would apply. For example, for the different combinations of verb phrase and adpositional phrase, a simplicity bias for consistent order predicts that combinations of phrases with harmonic orders should be preferred over those with non-harmonic orders, but makes no predictions on the learning/processing easiness of combinations of phrases with non-harmonic orders. The dependency-length minimization theory predicts that combinations of phrases with non-harmonic order in which the two heads are adjacent should be preferred over those in which the two heads are interrupted by the object nouns. For some word order patterns, however, the bias for harmonic order and the bias for dependency-length minimization make contrary predictions. For example, regarding the two complex phrases OVNPost and OVPrepN, the bias for harmonic order predicts that OVNPost should be preferred over OVPrepN because OVNPost has harmonic orders in the verb phrase and adpositional phrase whereas OVPrepN has non-harmonic orders in the two phrases; the bias for minimized dependency length, however, predicts that OVPrepN should be preferred over OVNPost because the distance between heads of the two phrases is shorter in OVPrepN than in OVNPost. Although there has been an artificial language learning experimental study testing the predictions of dependency-length minimization (Fedzechkina et al. 2018), the languages used in Fedzechkina et al. (2018) are always harmonic (i.e., either a head-initial language with VO and PrepN order or a head-final language with OV and NPost order). It is thus still unclear how these different biases interact and compete with each other during language learning and processing. Indeed, no artificial language learning experiment on harmony has yet investigated phrasal combinations.
4.2.5 Incorporating syntactic tests into artificial language learning experiments

Regarding word order of numeral classifiers, the set of experiments in Chapter 3 showed that learners had no preference for orders predicted either by the classifier-with-numeral hypothesis or the classifier-with-noun hypothesis. However, in almost all of the experiments, the results exhibit a bimodal distribution: some learners consistently chose one order and the others consistently chose another one. Very few learners showed substantial variation in the order they used. In Chapter 3, I discussed the phenomenon in English that in some contexts, combinations of numerals and mensural classifiers are likely to form a syntactic unit, whereas in some other contexts, mensural classifiers and nouns are likely to form a syntactic unit. I also discussed the possibility that the bimodal distribution of results might reflect learners inferring one of the two possible structures (e.g., based on their prior individual experience with mensural phrases), and then basing their order choices on that. However, the exact source of this across-participants variation remains unclear. In the experiments in Chapter 3, learners were only tested on the surface structure of the phrase consisting of a numeral, a classifier, and a noun. It is possible that a more refined grammatical judgment task would reveal how participants represent the structure by targeting specific contexts. One direction that could be explored in the future is thus going beyond the surface structure and integrating syntactic tests into artificial language learning experiments. For example, we could prompt learners to make grammaticality judgements on sentences derived from the two different structures in different contexts, such as asking learners to rate the grammaticality of coordinate structures in which either the combination of classifier and noun is omitted or only the noun is omitted. Data from such experiments may help uncover the source of variability among participants and provide more detailed evidence for the syntactic representation of [Num+Cl+N] phrase.
4.2.6 Order of numeral classifiers and gender: the influence of semantic contribution

In Chapter 3, I compared my experiments on numeral classifiers to those on gender in Saldana & Culbertson (2023). In my experiments on numeral classifiers, no reliable preference for any order between numeral, numeral classifier and noun was observed in almost all experiments. However, in Saldana & Culbertson (2023), learners exhibited a very strong bias to place gender morphemes closer to noun stems than number morphemes. I argued that the different results might be caused by the difference between the semantic contribution of gender morphemes to nouns in Saldana & Culbertson (2023) and the semantic contribution of sortal numeral classifiers to classifier phrases in my experiments. Specifically, gender morphemes in most experiments reported in Saldana & Culbertson (2023) have semantic contribution to the meanings of nouns whereas the sortal classifiers in my experiments do not have this kind of semantic contribution to the phrases. Future experiments could test whether a bias for the order between numeral, classifier, and noun emerges when sortal classifiers have some kind of semantic contribution to the phrase (i.e., the references of phrases differ based on the use of different sortal numeral classifiers, such as the Burmese examples mentioned in footnote 15 in Chapter 3), and whether learners still have a strong bias to place gender morphemes closer to noun stems than number morphemes when gender morphemes has no semantic contribution to the noun meanings but are just required by morphology.
Appendices
Appendix A

Appendix for Part I

A.1 Learning accuracy in non-critical testing phases

A.1.1 Learning accuracy in non-critical testing phases in Experiment 1

Participants’ productions in all non-critical testing phases were also coded for accuracy. For vocabulary accuracy across the experiment (i.e., accuracy for nouns, verbs, and adpositions in all non-critical testing phases), a production was coded as wrong if more than one segment of the word was wrong; otherwise the production was coded as correct. For verb phrase production trials, word order accuracy was also coded separately. For participants in the VO condition, VO order was coded as correct; for participants in the OV condition, OV order was coded as correct; otherwise the production was coded as incorrect (the coding criteria also applied to Experiment 2 and 3). Both vocabulary accuracy across all non-critical testing phases, and order accuracy in the verb phrase production trials were very high across populations and conditions, in all cases above 96% (recall that participants with accuracy below 75% on any one of these phases were excluded).

To ensure that differences in accuracy do not explain any differences across populations or conditions in the critical testing phase, we fitted Bayesian hierarchical logistic regression models with default priors for each non-critical testing phase.
The dependent variable was accuracy (i.e., vocabulary accuracy in noun testing, verb testing, adposition testing, and verb phrase production; response accuracy in verb phrase matching; word order accuracy in verb phrase production). The independent variables were always condition and population. Results of the models are shown in Table A.1 to Table A.6 below. A difference between conditions was supported by the model of the word order accuracy in verb phrase testing. Accuracy of the verb phrase order was likely to be lower in the OV condition compared to the VO condition (\( \hat{\beta} =-0.94, \text{CI} = [-2.24, -0.01] \)) (see Table A.5). However, accuracy in verb phrase order was very high in both conditions (VO condition: 99.71%, OV condition: 99.06%). Therefore, the difference in the strength of the bias for harmonic order between the two conditions is unlikely due to this difference, and certainly not due to insufficient learning of OV order. No other reliable differences in non-critical testing accuracy across the two conditions were found.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Est.Error</th>
<th>l-95% CI</th>
<th>u-95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.24</td>
<td>0.38</td>
<td>3.55</td>
<td>4.95</td>
</tr>
<tr>
<td>OV</td>
<td>-0.00</td>
<td>0.20</td>
<td>-0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>Mandarin speakers</td>
<td>-0.10</td>
<td>0.20</td>
<td>-0.49</td>
<td>0.29</td>
</tr>
<tr>
<td>OV: Mandarin speakers</td>
<td>-0.28</td>
<td>0.20</td>
<td>-0.67</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table A.1: Model parameter estimates, noun accuracy, Experiment 1.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Est.Error</th>
<th>l-95% CI</th>
<th>u-95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.88</td>
<td>0.60</td>
<td>3.72</td>
<td>5.95</td>
</tr>
<tr>
<td>OV</td>
<td>0.22</td>
<td>0.26</td>
<td>-0.28</td>
<td>0.75</td>
</tr>
<tr>
<td>Mandarin speakers</td>
<td>0.34</td>
<td>0.26</td>
<td>-0.16</td>
<td>0.84</td>
</tr>
<tr>
<td>OV: Mandarin speakers</td>
<td>-0.04</td>
<td>0.26</td>
<td>-0.53</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Table A.2: Model parameter estimates, verb accuracy, Experiment 1.
Predictor    Estimate    Est.Error    l-95% CI    u-95% CI
Intercept    5.16        0.47        4.34        6.16
OV           0.21        0.28        -0.33       0.76
Mandarin speakers 0.29      0.28        -0.25       0.85
OV: Mandarin speakers 0.07    0.28        -0.49       0.61

Table A.3: Model parameter estimates, adposition accuracy, Experiment 1.

Predictor    Estimate    Est.Error    l-95% CI    u-95% CI
Intercept    4.21        0.34        3.59        4.95
OV           -0.23       0.20        -0.63       0.16
Mandarin speakers -0.10      0.20        -0.50       0.28
OV: Mandarin speakers -0.21     0.20        -0.61       0.19

Table A.4: Model parameter estimates, vocab. accuracy of verb phrase production, Experiment 1.

Predictor    Estimate    Est.Error    l-95% CI    u-95% CI
Intercept    7.15        1.15        5.32        9.64
OV           -0.94       0.56        -2.24       -0.01
Mandarin speakers 0.22      0.51        -0.74       1.32
OV: Mandarin speakers 0.24    0.54        -0.74       1.39

Table A.5: Model parameter estimates, order accuracy of verb phrase production, Experiment 1.

Predictor    Estimate    Est.Error    l-95% CI    u-95% CI
Intercept    5.26        0.58        4.20        6.44
OV           0.05        0.26        -0.47       0.59
Mandarin speakers -0.07      0.26        -0.59       0.42
OV: Mandarin speakers 0.07    0.26        -0.45       0.57

Table A.6: Model parameter estimates, verb phrase matching accuracy, Experiment 1.

A.1.2 Learning accuracy in non-critical testing phases in Experiment 2

As in Experiment 1, we also analysed data in all non-critical testing phases on Experiment 2. We fitted a Bayesian hierarchical model for each phase (see Ap-
pendix A.1.1 for model details). Results of the models are shown in Table A.7 to Table A.12 below. A difference between conditions was supported by the model of the word order accuracy in verb phrase testing. Accuracy of the verb phrase order was likely to be lower in the OV condition compared to the VO condition ($\hat{\beta} = -1.17$, CI = [-2.50, -0.15]) (see Table A.11). However, the mean accuracy rates for verb phrase order are very high in both conditions (VO condition: 99.78%, OV condition: 98.76%), suggesting that both verb phrase orders were successfully acquired by learners. As for Experiment 1, the difference in the strength of the bias for harmonic order between the two conditions is unlikely due to this difference. No other reliable differences in non-critical testing accuracy across the two conditions were found.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Est.Error</th>
<th>l-95% CI</th>
<th>u-95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.53</td>
<td>0.53</td>
<td>3.55</td>
<td>5.42</td>
</tr>
<tr>
<td>OV</td>
<td>0.17</td>
<td>0.23</td>
<td>-0.30</td>
<td>0.62</td>
</tr>
<tr>
<td>NP</td>
<td>-0.02</td>
<td>0.23</td>
<td>-0.48</td>
<td>0.43</td>
</tr>
<tr>
<td>OV: NP</td>
<td>-0.14</td>
<td>0.24</td>
<td>-0.60</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Table A.7: Model parameter estimates, noun accuracy, Experiment 2.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Est.Error</th>
<th>l-95% CI</th>
<th>u-95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.52</td>
<td>0.80</td>
<td>2.79</td>
<td>5.89</td>
</tr>
<tr>
<td>OV</td>
<td>0.17</td>
<td>0.24</td>
<td>-0.30</td>
<td>0.63</td>
</tr>
<tr>
<td>NP</td>
<td>0.01</td>
<td>0.24</td>
<td>-0.45</td>
<td>0.48</td>
</tr>
<tr>
<td>OV: NP</td>
<td>-0.11</td>
<td>0.24</td>
<td>-0.61</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Table A.8: Model parameter estimates, verb accuracy, Experiment 2.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Est.Error</th>
<th>l-95% CI</th>
<th>u-95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.21</td>
<td>0.48</td>
<td>4.34</td>
<td>6.26</td>
</tr>
<tr>
<td>OV</td>
<td>0.12</td>
<td>0.29</td>
<td>-0.45</td>
<td>0.72</td>
</tr>
<tr>
<td>NP</td>
<td>0.19</td>
<td>0.29</td>
<td>-0.36</td>
<td>0.77</td>
</tr>
<tr>
<td>OV: NP</td>
<td>-0.04</td>
<td>0.29</td>
<td>-0.61</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Table A.9: Model parameter estimates, adposition/adjective accuracy, Experiment 2.
## A.1.3 Learning accuracy in non-critical testing phases in Experiment 3

As in Experiments 1 and 2, we also analyzed data in all non-critical testing phases of Experiment 3. We fitted a Bayesian hierarchical model for each phase (see Appendix A.1.1 for model details. One difference is that here the extrapolation phrase type predictor was Helmert coded). Results of the models are shown in Table A.13 to Table A.18. A difference between conditions was found in accuracy of noun production in the noun testing phase: accuracy in the OV condition was likely to be slightly higher than that in the VO condition ($\hat{\beta} = 0.23, CI = [0.05, 0.43]$) (see Table A.13). However, the mean accuracy rates for noun testing in the two conditions

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Est.Error</th>
<th>l-95% CI</th>
<th>u-95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.24</td>
<td>0.35</td>
<td>3.60</td>
<td>4.98</td>
</tr>
<tr>
<td>OV</td>
<td>0.06</td>
<td>0.21</td>
<td>-0.35</td>
<td>0.48</td>
</tr>
<tr>
<td>NP</td>
<td>-0.17</td>
<td>0.21</td>
<td>-0.58</td>
<td>0.24</td>
</tr>
<tr>
<td>OV: NP</td>
<td>0.07</td>
<td>0.21</td>
<td>-0.35</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Table A.10: Model parameter estimates, vocab. accuracy of verb phrase production, Experiment 2.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Est.Error</th>
<th>l-95% CI</th>
<th>u-95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>7.49</td>
<td>1.20</td>
<td>5.71</td>
<td>10.25</td>
</tr>
<tr>
<td>OV</td>
<td>-1.17</td>
<td>0.60</td>
<td>-2.50</td>
<td>-0.15</td>
</tr>
<tr>
<td>NP</td>
<td>0.32</td>
<td>0.58</td>
<td>-0.77</td>
<td>1.52</td>
</tr>
<tr>
<td>OV: NP</td>
<td>0.03</td>
<td>0.57</td>
<td>-1.10</td>
<td>1.18</td>
</tr>
</tbody>
</table>

Table A.11: Model parameter estimates, order accuracy of verb phrase production, Experiment 2.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Est.Error</th>
<th>l-95% CI</th>
<th>u-95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.42</td>
<td>0.73</td>
<td>4.01</td>
<td>6.86</td>
</tr>
<tr>
<td>OV</td>
<td>-0.05</td>
<td>0.27</td>
<td>-0.58</td>
<td>0.50</td>
</tr>
<tr>
<td>NP</td>
<td>0.12</td>
<td>0.28</td>
<td>-0.43</td>
<td>0.67</td>
</tr>
<tr>
<td>OV: NP</td>
<td>-0.03</td>
<td>0.28</td>
<td>-0.60</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Table A.12: Model parameter estimates, verb phrase matching accuracy, Experiment 2.
are generally very high (VO condition: 93.96%; OV condition: 95.94%), suggesting that nouns in the two conditions were successfully acquired by learners. Any differences in the probability of producing harmonic order across the two conditions is thus unlikely to be driven by this. No other reliable differences in accuracy were found.

Table A.13: Model parameter estimates, noun accuracy, Experiment 3.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Est.Error</th>
<th>l-95% CI</th>
<th>u-95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.61</td>
<td>0.33</td>
<td>2.95</td>
<td>4.20</td>
</tr>
<tr>
<td>Condition.OVMintercept</td>
<td>0.23</td>
<td>0.10</td>
<td>0.05</td>
<td>0.43</td>
</tr>
<tr>
<td>PhraseType.ActiveAdjMStativeAdj</td>
<td>0.08</td>
<td>0.11</td>
<td>-0.15</td>
<td>0.29</td>
</tr>
<tr>
<td>PhraseType.PPMStativeAdj.ActiveAdj</td>
<td>0.07</td>
<td>0.07</td>
<td>-0.07</td>
<td>0.22</td>
</tr>
<tr>
<td>Condition.OVMintercept: PhraseType.ActiveAdjMStativeAdj</td>
<td>-0.03</td>
<td>0.11</td>
<td>-0.25</td>
<td>0.19</td>
</tr>
<tr>
<td>Condition.OVMintercept: PhraseType.PPMStativeAdj.ActiveAdj</td>
<td>-0.06</td>
<td>0.07</td>
<td>-0.21</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Table A.14: Model parameter estimates, verb accuracy, Experiment 3.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Est.Error</th>
<th>l-95% CI</th>
<th>u-95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.11</td>
<td>0.64</td>
<td>2.63</td>
<td>5.14</td>
</tr>
<tr>
<td>Condition.OVMintercept</td>
<td>0.06</td>
<td>0.11</td>
<td>-0.15</td>
<td>0.28</td>
</tr>
<tr>
<td>PhraseType.ActiveAdjMStativeAdj</td>
<td>-0.08</td>
<td>0.12</td>
<td>-0.33</td>
<td>0.16</td>
</tr>
<tr>
<td>PhraseType.PPMStativeAdj.ActiveAdj</td>
<td>0.03</td>
<td>0.08</td>
<td>-0.13</td>
<td>0.19</td>
</tr>
<tr>
<td>Condition.OVMintercept: PhraseType.ActiveAdjMStativeAdj</td>
<td>0.02</td>
<td>0.13</td>
<td>-0.23</td>
<td>0.27</td>
</tr>
<tr>
<td>Condition.OVMintercept: PhraseType.PPMStativeAdj.ActiveAdj</td>
<td>-0.12</td>
<td>0.09</td>
<td>-0.28</td>
<td>0.05</td>
</tr>
<tr>
<td>Predictor</td>
<td>Estimate</td>
<td>Est.Error</td>
<td>l-95% CI</td>
<td>u-95% CI</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>----------</td>
<td>-----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>Intercept</td>
<td>4.81</td>
<td>0.44</td>
<td>3.88</td>
<td>5.54</td>
</tr>
<tr>
<td>Condition.OVMintercept</td>
<td>-0.11</td>
<td>0.15</td>
<td>-0.40</td>
<td>0.18</td>
</tr>
<tr>
<td>PhraseType.ActiveAdjMStativeAdj</td>
<td>-0.07</td>
<td>0.17</td>
<td>-0.42</td>
<td>0.25</td>
</tr>
<tr>
<td>PhraseType.PPMStativeAdj.ActiveAdj</td>
<td>-0.03</td>
<td>0.11</td>
<td>-0.24</td>
<td>0.19</td>
</tr>
<tr>
<td>Condition.OVMintercept:</td>
<td>-0.20</td>
<td>0.17</td>
<td>-0.54</td>
<td>0.14</td>
</tr>
<tr>
<td>PhraseType.ActiveAdjMStativeAdj</td>
<td>-0.05</td>
<td>0.11</td>
<td>-0.27</td>
<td>0.16</td>
</tr>
<tr>
<td>PhraseType.PPMStativeAdj.ActiveAdj</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.15: Model parameter estimates, adposition/adjective accuracy, Experiment 3.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Est.Error</th>
<th>l-95% CI</th>
<th>u-95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.60</td>
<td>0.28</td>
<td>3.03</td>
<td>4.10</td>
</tr>
<tr>
<td>Condition.OVMintercept</td>
<td>0.08</td>
<td>0.09</td>
<td>-0.10</td>
<td>0.25</td>
</tr>
<tr>
<td>PhraseType.ActiveAdjMStativeAdj</td>
<td>0.07</td>
<td>0.10</td>
<td>-0.14</td>
<td>0.27</td>
</tr>
<tr>
<td>PhraseType.PPMStativeAdj.ActiveAdj</td>
<td>-0.03</td>
<td>0.07</td>
<td>-0.16</td>
<td>0.10</td>
</tr>
<tr>
<td>Condition.OVMintercept:</td>
<td>-0.05</td>
<td>0.10</td>
<td>-0.26</td>
<td>0.15</td>
</tr>
<tr>
<td>PhraseType.ActiveAdjMStativeAdj</td>
<td>-0.08</td>
<td>0.06</td>
<td>-0.21</td>
<td>0.05</td>
</tr>
<tr>
<td>PhraseType.PPMStativeAdj.ActiveAdj</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.16: Model parameter estimates, vocab. accuracy of verb phrase production, Experiment 3.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Est.Error</th>
<th>l-95% CI</th>
<th>u-95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.87</td>
<td>0.34</td>
<td>5.27</td>
<td>6.54</td>
</tr>
<tr>
<td>Condition.OVMintercept</td>
<td>-0.26</td>
<td>0.19</td>
<td>-0.63</td>
<td>0.09</td>
</tr>
<tr>
<td>PhraseType.ActiveAdjMStativeAdj</td>
<td>0.16</td>
<td>0.19</td>
<td>-0.20</td>
<td>0.54</td>
</tr>
<tr>
<td>PhraseType.PPMStativeAdj.ActiveAdj</td>
<td>0.12</td>
<td>0.15</td>
<td>-0.15</td>
<td>0.43</td>
</tr>
<tr>
<td>Condition.OVMintercept:</td>
<td>0.13</td>
<td>0.19</td>
<td>-0.24</td>
<td>0.51</td>
</tr>
<tr>
<td>PhraseType.ActiveAdjMStativeAdj</td>
<td>-0.21</td>
<td>0.15</td>
<td>-0.51</td>
<td>0.06</td>
</tr>
<tr>
<td>PhraseType.PPMStativeAdj.ActiveAdj</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.17: Model parameter estimates, order accuracy of verb phrase production, Experiment 3.

152
<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Est.Error</th>
<th>l-95% CI</th>
<th>u-95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.72</td>
<td>0.22</td>
<td>4.26</td>
<td>5.14</td>
</tr>
<tr>
<td>Condition.OVMintercept</td>
<td>0.16</td>
<td>0.11</td>
<td>-0.06</td>
<td>0.38</td>
</tr>
<tr>
<td>PhraseType.ActiveAdjMStativeAdj</td>
<td>-0.01</td>
<td>0.13</td>
<td>-0.27</td>
<td>0.24</td>
</tr>
<tr>
<td>PhraseType.PPMStativeAdj.ActiveAdj</td>
<td>-0.10</td>
<td>0.08</td>
<td>-0.25</td>
<td>0.06</td>
</tr>
<tr>
<td>Condition.OVMintercept:PhraseType.ActiveAdjMStativeAdj</td>
<td>0.03</td>
<td>0.13</td>
<td>-0.23</td>
<td>0.28</td>
</tr>
<tr>
<td>Condition.OVMintercept:PhraseType.PPMStativeAdj.ActiveAdj</td>
<td>-0.08</td>
<td>0.08</td>
<td>-0.24</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table A.18: Model parameter estimates, accuracy of verb phrase matching, Experiment 3.

### A.2 Second language experience in Experiments 2 and 3

As reported in the main text, we investigated the potential effect of participants’ self-reported language experience on order in Experiments 2 and 3. In Experiment 2, 27 participants reported speaking a second language with a different order from English. Word order in the verb phrase and the noun phrase with adjective in these languages were: VO and NAdj, OV and AdjN, and no dominant order in the verb phrase and AdjN. We fitted a Bayesian hierarchical logistic regression model to evaluate whether the probability of producing harmonic order in the noun phrase is predicted by the word order type of participant’s second language, verb phrase order the participant was trained on (both sum-coded), and their interaction. We experimented with different sets of priors, but unfortunately the model failed to converge in each case. As shown in Figure A.1, there are not any obvious differences based on language experience.

In the two noun phrase extrapolation conditions of Experiment 3, 84 participants reported speaking a second language with a different order from English: VO and NAdj, VO and no dominant order between adjective and noun, OV and AdjN, no dominant order in the verb phrase and AdjN. We fitted a Bayesian hierarchical logistic regression model testing whether harmonic order in the noun phrase is predicted by second language order, and verb phrase training conditions
Figure A.1: Proportion of responses using harmonic order across verb phrase training conditions and order types of learners’ second language in Experiment 2; error bars indicate 95% confidence intervals on the mean of by-participant means; colored points indicate individual participant means; the dashed line shows chance level (50%). Proportion harmonic responses of learners with different second language experience in the two conditions are quiet similar, suggesting that the difference in producing harmonic order in the two conditions in Experiment 2 is unlikely to be due to learners’ second language experience.

(both predictors sum-coded), and their interaction. The model with the default prior parameter values specified in brms package did not converge. We therefore adjusted the priors to attain model convergence. We used priors centered on zero to reflect that we have no clear expectations about the parameter values. In particular, the priors we used for the intercept and the other fixed effects (including the interaction term) were all Normal(0, 5). These priors permit a wide range of values, both positive and negative. Four sampling chains ran for 3000 iterations with a 1500 warm-up iterations for each chain, yielding 6000 samples for each parameter in the posterior distribution. The posterior distribution indicated no reliable effects of second language experience across training conditions.

In the adpositional phrase extrapolation condition of Experiment 3, 7 participants reported speaking a second language that has different order from English. We fitted a Bayesian hierarchical logistic regression model testing whether harmonic order in the noun phrase is predicted by second language order, and verb
phrase training conditions (both predictors sum coded), and their interaction. We experimented with different sets of priors, but unfortunately the model failed to converge in each case. As shown in Figure A.2, 6 participants produced harmonic order regardless of the order type of their second languages. Only 1 participant who spoke a second language with OV and NPost order in the OV condition produced non-harmonic order. This suggests that the strong bias for harmony we observed in the adpositional phrase extrapolation condition is unlikely to be due to the influence of learners’ second language experience.

Figure A.2: Proportion of responses using harmonic order across verb phrase training conditions and order types of learners’ second language in the adpositional phrase extrapolation condition in Experiment 3; colored points indicate individual participant means; the dashed line shows chance level (50%). 6 of 7 participants produced harmonic order regardless of the order type of their second languages, suggesting that the strong bias for harmony we observed in the adpositional phrase extrapolation condition in Experiment 3 is unlikely to be due to the influence of learners’ second language experience.
A.3 Learners’ responses in translating artificial words/phrases into their native languages

In Experiments 1 and 2, we prompted learners to translate the artificial adpositions and adjectives into their native language. In Experiment 3, we asked learners to translate both the artificial words and the artificial phrases at the critical test into their native language. To avoid redundancy, here we reported their responses for translating artificial words in Experiments 1 and 2, and their responses in translating artificial phrases in Experiment 3, which can directly indicate whether learners acquired the artificial phrases as noun phrases or adpositional phrases (see Table A.19 to Table A.25).

<table>
<thead>
<tr>
<th>Meaning</th>
<th>Learners’ translation</th>
<th>VO condition</th>
<th>OV condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘left of’</td>
<td>Preposition (e.g., <em>to left, left of</em>)</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td><em>left</em></td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>‘right of’</td>
<td>Preposition (e.g., <em>to right, right of</em>)</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td><em>right</em></td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>‘in front of’</td>
<td>Preposition (e.g., <em>in front of</em>)</td>
<td>21</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td><em>front</em></td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>‘behind’</td>
<td>Preposition (e.g., <em>behind</em>)</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td><em>back</em></td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A.19: Counts of different types of translations of artificial adpositions by English speakers, Experiment 1.
<table>
<thead>
<tr>
<th>Meaning</th>
<th>Learners’ translation</th>
<th>VO condition</th>
<th>OV condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘left of’</td>
<td>zai+...+Postposition/Noun (e.g., zai...zuo ‘at...left’, zai...zuobian ‘at...left.side’)</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Postposition/Noun (e.g., zuo(bian) ‘left(.side)’)</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Verb phrase (i.e., zhan wuti zuoce ‘stand object left.side’)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>‘right of’</td>
<td>zai+...+Postposition/Noun (e.g., zai...you ‘at...right’, zai...youbian ‘at...left.side’)</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Postposition/Noun (e.g., you(bian) ‘right(.side)’)</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Verb phrase (i.e., zhan wuti youce ‘stand object right.side’)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>‘in front of’</td>
<td>zai+...+Postposition/Noun (e.g., zai...qian ‘at...front’, zai...qianbian ‘at...front.side’)</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Postposition/Noun (e.g., qian(bian) ‘front(.side)’)</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Verb phrase (i.e., zhan wuti qiance ‘stand object front.side’)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>‘behind’</td>
<td>zai+...+Postposition/Noun (e.g., zai...hou ‘at...back’, zai...houbian ‘at...back side’)</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Postposition/Noun (e.g., houbian) ‘back(.side)’)</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Verb phrase (i.e., zhan wuti houce ‘stand object back.side’)</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A.20: Counts of different types of translations of artificial adpositions by Mandarin speakers, Experiment 1.
<table>
<thead>
<tr>
<th>Meaning</th>
<th>Learners’ translation</th>
<th>VO condition</th>
<th>OV condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘red’</td>
<td>red</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>‘green’</td>
<td>green</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>‘blue’</td>
<td>blue</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>‘yellow’</td>
<td>yellow</td>
<td>20</td>
<td>23</td>
</tr>
</tbody>
</table>

Table A.21: Counts of different types of translations of artificial adjectives by English speakers, Experiment 2.

<table>
<thead>
<tr>
<th>Meaning</th>
<th>Learners’ translation</th>
<th>VO condition</th>
<th>OV condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘red object’</td>
<td>red+object (e.g., red fugan)</td>
<td>39</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>red</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>‘green object’</td>
<td>green+object (e.g., green fugan)</td>
<td>39</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>green</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>‘blue object’</td>
<td>blue+object (e.g., blue fugan)</td>
<td>39</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>blue</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>‘yellow object’</td>
<td>yellow+object (e.g., yellow fugan)</td>
<td>39</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>yellow</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

Table A.22: Counts of different types of translations of noun+color adjective phrases at the critical test, Experiment 3.
<table>
<thead>
<tr>
<th>Meaning</th>
<th>Learners’ translation</th>
<th>VO condition</th>
<th>OV condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘spotted object’</td>
<td>Adj(+object) (e.g., spotted fugan, spotty fugan)</td>
<td>38</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Adj+pattern+object (e.g., pokadot pattern teacup)</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Other (e.g., fugan with polka dot pattern)</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>‘stripped object’</td>
<td>Adj(+object) (e.g., striped fugan, stripy fugan)</td>
<td>38</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Other (e.g., fugan with diagonal stripes)</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>‘checkered object’</td>
<td>Adj(+object) (e.g., checkered fugan, houndstooth fugan)</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Adj+pattern(+object) (e.g., plaid pattern teacup)</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Noun+object (e.g., pattern fugan)</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Other (e.g., fugan with checkered pattern)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>‘woven object’</td>
<td>Adj(+object) (e.g., woven fugan, patterned fugan)</td>
<td>32</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Adj+pattern(+object) (e.g., overlapping pattern teacup)</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>pattern(+object) (e.g., pattern, pattern fugan)</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Other (i.e., fugan with weave/crosshatch pattern)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

Table A.23: Counts of different types of translations of noun+texture adjective phrases at the critical test, Experiment 3.
<table>
<thead>
<tr>
<th>Meaning</th>
<th>Learners’ translation</th>
<th>VO condition</th>
<th>OV condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘broken object’</td>
<td>Adj(+object)</td>
<td>39</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>(e.g., broken fugan)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Verb(+object)</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>(e.g., broke fugan, break)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sentence</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(i.e., the fugan is broken)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>‘twisted object’</td>
<td>Adj(+object)</td>
<td>38</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>(e.g., twisted fugan, squeezed fugan)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Verb(+object)</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(e.g., twist fugan)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sentence (i.e., the fugan is twisted)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>‘squashed object’</td>
<td>Adj(+object)</td>
<td>38</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>(e.g., squashed fugan, flatten fugan)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Verb(+object)</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(i.e., squash fugan, stretch)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sentence</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(i.e., the fugan has been squashed)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>‘flipped object’</td>
<td>Adj(+object)</td>
<td>39</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>(e.g., flipped fugan, upside down fugan)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Verb(+object)</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(e.g., flip fugan)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sentence</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(i.e., the fugan is upside down)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Table A.24: Counts of different types of translations of noun+active adjective phrases at the critical test, Experiment 3.
<table>
<thead>
<tr>
<th>Meaning</th>
<th>Learners’ translation</th>
<th>VO condition</th>
<th>OV condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘in front of object’</td>
<td>Preposition(+object) (e.g., in front of fugan, infront fugan)</td>
<td>34</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Verb phrase (e.g., standing in front of the bowl)</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>‘behind object’</td>
<td>Preposition(+object) (e.g., behind fugan)</td>
<td>35</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Verb phrase (e.g., standing behind the bowl)</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Other (i.e., back)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>‘on the left of object’</td>
<td>Preposition(+object) (e.g., to the left of fugan, on the left of fugan)</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>left</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Verb phrase (e.g., standing left side of the bowl)</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>‘on the right of object’</td>
<td>Preposition(+object) (e.g., to the right of fugan, on the right of fugan)</td>
<td>29</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>right</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Verb phrase (e.g., standing right side of the bowl)</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A.25: Counts of different types of translations of adpositional phrases at the critical test, Experiment 3.
A.4 The Bayesian model with weakly informative priors in Experiment 3

To examine the robustness of our findings in Experiment 3, we additionally run a Bayesian hierarchical logistic regression model with informative priors based on the findings of Experiments 1 and 2. The structure of the model is the same as that detailed in Experiment 3, differing solely in the priors.

In the current model, the prior we used for the intercept, which represents the grand mean, was Normal(2, 2.5). This prior suggests that we overall expect a harmony bias but still permit non-harmonic orders. The prior we used for the coefficient for condition was Normal(-1, 2), suggesting that we expect that the harmony bias is weaker in the OV condition than in the VO condition, but we still permit it to be stronger. This is based on the findings that harmony bias for both adpositional phrase and noun phrase is weaker in the OV condition than in the VO condition in Experiments 1 and 2. The priors we used for the coefficients for active adjectives versus stative adjectives and that for the interaction of condition and active adjective versus stative adjectives were both Normal(0, 5), which reflect that we have no clear expectations about these two parameter values. The prior for the coefficient for adpositional phrases versus noun phrase (including active and stative adjectives) was Normal(1, 2), suggesting that we believe the harmony bias in adpositional phrase is likely to be stronger than in noun phrases (including active and stative adjectives) but we still permit it to be weaker. This is based on the results that the harmony bias in adpositional phrase in Experiment 1 is quite strong but no such bias in the noun phrase with color adjectives was observed in Experiment 2. The prior for the coefficient for interaction of condition and adpositional phrase versus noun phrase (including active and stative adjectives) was Normal(1, 2). This suggests our expectation that the impact of phrase type is likely to become more pronounced when moving from the VO condition to the OV condition, while still allowing for the possibility of it diminishing.

Table A.26 shows the parameter estimates of this model. The conclusions drawn from this model align with those from the model with default priors.
<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>Est.Error</th>
<th>l-95% CI</th>
<th>u-95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9.04</td>
<td>0.98</td>
<td>7.19</td>
<td>11.03</td>
</tr>
<tr>
<td>Condition.OVMintercept</td>
<td>-2.94</td>
<td>0.79</td>
<td>-4.54</td>
<td>-1.43</td>
</tr>
<tr>
<td>PhraseType.ActiveAdjMStativeAdj</td>
<td>1.66</td>
<td>1.81</td>
<td>-1.90</td>
<td>5.31</td>
</tr>
<tr>
<td>PhraseType.PPMStativeAdj.ActiveAdj</td>
<td>3.12</td>
<td>1.43</td>
<td>0.33</td>
<td>5.92</td>
</tr>
<tr>
<td>Condition.OVMintercept:</td>
<td>4.88</td>
<td>1.88</td>
<td>1.33</td>
<td>8.70</td>
</tr>
<tr>
<td>PhraseType.ActiveAdjMStativeAdj</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition.OVMintercept:</td>
<td>1.08</td>
<td>1.43</td>
<td>-1.72</td>
<td>3.90</td>
</tr>
<tr>
<td>PhraseType.PPMStativeAdj.ActiveAdj</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.26: Parameter estimates in the additional model with informative priors in Experiment 3.
Appendix B

Appendix for Part II

B.1 Vocabulary accuracy in non-critical testing phases

B.1.1 Vocabulary accuracy in non-critical testing phases in Experiment 1A and 1B

Participants’ selections in all non-critical testing phases in Experiments 1A and 1B were coded for accuracy. For each syntactic category, the accuracy data includes both the accuracy in the separate testing phase for that syntactic category and the accuracy of that syntactic category in the mixed testing phase. The mean accuracy rates of all syntactic categories are very high in both Experiment 1A (above 97%) and 1B (above 98%) (recall that participants whose accuracy rate was below 85% in any of the non-critical testing phases were excluded from data analysis). The results of vocabulary accuracy in Experiments 1A and 1B are shown in Figure B.1 and Figure B.2, respectively.

B.1.2 Vocabulary accuracy in non-critical testing phases in Experiment 2

Participants’ productions in all non-critical testing phases in Experiment 2 were also coded for accuracy. A production was coded as wrong if more than one seg-
Figure B.1: Vocabulary accuracy of nouns, numerals, and classifiers in Experiment 1A; error bars indicate 94% confidence intervals on the mean of by-participant means.

Figure B.2: Vocabulary accuracy of nouns, numerals, and classifiers in Experiment 1B; error bars indicate 95% confidence intervals on the mean of by-participant means.
Figure B.3: Vocabulary accuracy of nouns, numerals, and classifiers in Experiment 2; error bars indicate 95% confidence intervals on the mean of by-participant means.

ment of the word was wrong; otherwise the production was coded as correct. As in Experiments 1A and 1B, the accuracy data includes both the accuracy in the separate testing phase for that syntactic category and the accuracy of that syntactic category in the mixed testing phase. The mean accuracy rates of all syntactic categories in Experiment 2 are very high (above 94%) (recall that participants with an accuracy rate below 75% in any of the non-critical testing phases were excluded from data analysis). The results of vocabulary accuracy in Experiment 2 are shown in Figure B.3.

B.1.3 Vocabulary accuracy in non-critical testing phases in Experiment 3

As in Experiment 2, we also coded the vocabulary accuracy in all non-critical testing phases in Experiment 3. The mean accuracy rates of all syntactic categories in the mixed testing phase are also very high (above 95%) (recall that participants with an accuracy rate below 75% for any syntactic category in the mixed testing phase were excluded from data analysis). The results of vocabulary accuracy in non-critical
Figure B.4: Vocabulary accuracy of nouns, numerals, and classifiers in each non-critical testing phase in Experiment 3; error bars indicate 95% confidence intervals on the mean of by-participant means.

testing phases in Experiment 3 are shown in Figure B.4.
To check whether learners correctly acquired the artificial words, we analyzed learners’ responses for translating artificial words into their native languages in each experiment, which are shown in Table B.1 to Table B.4. Learners’ responses suggest that in most cases, they did acquire the meanings of artificial words in the experiments.

<table>
<thead>
<tr>
<th>Syntactic category</th>
<th>Meaning</th>
<th>Learners’ translation</th>
<th>Condition 1</th>
<th>Condition 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sortal classifier</td>
<td>Human</td>
<td><em>human, person, people</em></td>
<td>27</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Other (e.g., <em>there are, growing up</em>)</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Non-human</td>
<td>non-human, nature, things, objects...</td>
<td>24</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other (e.g., <em>outside, mammal</em>)</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Noun</td>
<td>‘flower’</td>
<td><em>flower, plant, object</em></td>
<td>29</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Other (e.g., <em>3</em>)</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>‘teenager’</td>
<td><em>boy, person, man, girl</em></td>
<td>29</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Other (i.e., <em>2</em>)</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Numeral</td>
<td>‘two’</td>
<td><em>two</em></td>
<td>28</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Other (i.e., <em>3</em>)</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>‘three’</td>
<td><em>three</em></td>
<td>27</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Other (i.e., <em>1, many</em>)</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No response</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Table B.1: Counts of different types of translations of artificial words by learners, Experiment 1A.
### Table B.2: Counts of different types of translations of artificial words by learners, Experiment 1B.

<table>
<thead>
<tr>
<th>Syntactic category</th>
<th>Meaning</th>
<th>Learners’ translation</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sortal classifier</td>
<td>Human</td>
<td>human, people, person...</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other (e.g., there are, growing)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No response</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Non-human</td>
<td>non-human, nature, things, objects...</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other (e.g., there are)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No response</td>
<td>3</td>
</tr>
<tr>
<td>Noun</td>
<td>‘flower’</td>
<td>flower, plant, rose, object...</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>‘teenager’</td>
<td>boy, man, person, girl...</td>
<td>30</td>
</tr>
<tr>
<td>Numeral</td>
<td>‘two’</td>
<td>two</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>‘three’</td>
<td>three</td>
<td>30</td>
</tr>
</tbody>
</table>

### Table B.3: Counts of different types of translations of artificial words by learners, Experiment 2.

<table>
<thead>
<tr>
<th>Syntactic category</th>
<th>Meaning</th>
<th>Learners’ translation</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sortal classifier</td>
<td>Human</td>
<td>human, people, person...</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other (e.g., group, of them)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No response</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Non-human</td>
<td>non-human/nature/ things/objects...</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other (e.g., group)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No response</td>
<td>3</td>
</tr>
<tr>
<td>Noun</td>
<td>‘flower’</td>
<td>flower/plant/rose/vase...</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>‘teenager’</td>
<td>boy/man/person/girl...</td>
<td>27</td>
</tr>
<tr>
<td>Numeral</td>
<td>‘two’</td>
<td>two</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>‘three’</td>
<td>three</td>
<td>29</td>
</tr>
</tbody>
</table>

169
<table>
<thead>
<tr>
<th>Syntactic category</th>
<th>Meaning</th>
<th>Learners’ translation</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mensural classifier</td>
<td>'whole'</td>
<td>whole/complete...</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other (i.e., big)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>'piece'</td>
<td>piece/slice/part</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other (i.e., small, cut)</td>
<td>2</td>
</tr>
<tr>
<td>Noun</td>
<td>'pizza'</td>
<td>pizza</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>'cucumber'</td>
<td>cucumber</td>
<td>30</td>
</tr>
<tr>
<td>Numeral</td>
<td>'two'</td>
<td>two</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>'three'</td>
<td>three</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other (i.e., more)</td>
<td>1</td>
</tr>
</tbody>
</table>

Table B.4: Counts of different types of translations of artificial words by learners, Experiment 3.
Appendix C

Appendix for instructions for all experiments

C.1 Instructions for experiments in Part I

C.1.1 Instructions for Experiment 1

For English speakers:

Welcome text: Welcome! You are going to learn a new language. Your participation will help us to understand how English speakers learn and use this language. Instructions will be provided at the beginning of each section. Please read and follow the instructions carefully. Now please press spacebar to start.

Noun training phase: In each trial, you will see a picture of an object and hear the name of the object. Repeat what you hear, then press the spacebar to continue.

Instruction for each trial: Repeat what you hear, then press the spacebar to continue.

Noun testing phase: In each trial, you will see a picture of an object you previously saw. Say its name, and press the spacebar, then you will hear the correct answer.

Instruction for each trial: Say its name, then press the spacebar to continue.
**Verb training phase:** In each trial, you will see a picture of a boy and hear a word to describe what the boy is doing. Repeat what you hear, then press the spacebar to continue.

*Instruction for each trial:* Repeat what you hear, then press the spacebar to continue.

**Verb testing phase:** In each trial, you will see a picture of the boy you previously saw. Say what the boy is doing, and press the spacebar, then you will hear the correct answer.

*Instruction for each trial:* Say what the boy is doing, then press the spacebar to continue.

**Verb phrase training phase:** In each trial, you will see a picture of a boy and an object. You will also hear an utterance to describe what the boy is doing. Repeat what you hear, then press the spacebar to continue.

*Instruction for each trial:* Repeat what you hear, then press the spacebar to continue.

**Verb phrase matching phase:** In each trial, you will see four pictures and hear an utterance to describe what the boy is doing in one of the pictures. Click on the picture that matches the description, then you will see the correct picture.

*Instruction for each trial:* Click on the picture that matches the description.

**Verb phrase production testing phase:** In each trial, you will see a picture of a boy and an object you previously saw. Describe what is happening in the picture, and press the spacebar, then you will hear the correct answer.

*Instruction for each trial:* Describe the full picture, then press the spacebar to continue.

**Adposition training phase:** In each trial, you will see four pictures. In each picture, the boy is in the same position relative to an object. You will hear a word to describe where the boy is. Repeat what you hear, then press the spacebar to continue.

*Instruction for each trial:* Repeat what you hear, then press the spacebar to continue.
Adposition testing phase: In each trial, you will see four pictures as before. Describe where the boy is in the pictures, and press the spacebar, then you will hear the correct answer.

Instruction for each trial: Describe where the boy is, then press the spacebar to continue.

Adpositional phrase testing phase: This is your last test. In each trial, you will see a picture of a boy and an object. Describe the full picture. You will need to use two words. You have not heard descriptions like this before, so just do your best. When you are finished, press the spacebar to continue.

Instruction for each trial: Describe the full picture, then press spacebar to continue.

For Mandarin speakers:

Welcome text: 您好！欢迎您参加本实验。在本实验中，您将学习一门新的语言，您的参与将有助于我们理解汉语母语者如何学习和使用这门语言。实验中，我们会提供具体的指导语。请您认真阅读并按照指导语的要求完成实验。现在请您按空格键，开始实验。

Noun training phase: 屏幕上每次会呈现一张物体的图片，同时您会听到该物体的名称。请重复您所听到的词语，然后按空格键继续。

Instruction for each trial: 请重复您所听到的词语，然后按空格键继续。

Noun testing phase: 屏幕上每次会呈现一张物体的图片，该图片您之前看过。请说出该物体的名称，然后按空格键，之后您会听到正确答案。

Instruction for each trial: 请说出该物体的名称，然后按空格键继续。

Verb training phase: 屏幕上每次会呈现一张男孩的图片，同时您会听到一个词语来描述这个男孩在做什么。请重复您所听到的词语，然后按空格键继续。

Instruction for each trial: 请重复您所听到的词语，然后按空格键继续。

Verb testing phase: 屏幕上每次会呈现一张男孩的图片，该图片您之前看过。请说出这个男孩在做什么，然后按空格键，之后您会听到正确答案。

Instruction for each trial: 请说出这个男孩在做什么，然后按空格键继续。
Verb phrase training phase: 屏幕上每次会呈现一张图片，该图片上有一个男孩和一个物体。同时您会听到一句话语，该话语描述这个男孩在做什么。请重复您所听到的话语，然后按空格键继续。

Instruction for each trial: 请重复您所听到的话语，然后按空格键继续。

Verb phrase matching phase: 屏幕上每次会呈现四张图片，同时您会听到一句话语，该话语描述其中一幅图片中的男孩在做什么。请您点击鼠标选出与该描述相符的图片，然后按空格键，之后您会看到正确答案。

Instruction for each trial: 请点击鼠标选择与话语描述相符的图片。

Verb phrase production testing phase: 屏幕上每次会呈现一张男孩和物体的图片，该图片您之前看过。请描述图片中发生了什么，然后按空格键，之后您会听到正确答案。

Instruction for each trial: 请完整描述图片，然后按空格键继续。

Adposition training phase: 屏幕上每次会呈现四张图片。每张图片中，男孩都位于相对于物体的同一位置。您会听到一个词语，该词语描述这个男孩在哪。请重复您所听到的词语，然后按空格键继续。

Instruction for each trial: 请重复您所听到的词语，然后按空格键继续。

Adposition testing phase: 如同上一部分，屏幕上每次会呈现四张图片。请描述这些图片中的男孩在哪，然后按空格键，之后您会听到正确答案。

Instruction for each trial: 请描述这个男孩在哪，然后按空格键继续。

Adpositional phrase testing phase: 这是最后一部分测试。屏幕上每次会呈现一张男孩和物体的图片。请您完整描述该图片。您需要使用两个词语，您之前没有听到过类似的描述，请您按照您的想法来描述。说出话语后，请按空格键继续。

Instruction for each trial: 请完整描述图片，然后按空格键继续。

C.1.2 Instructions for Experiment 2

The welcome text and the instructions for the noun training and testing, verb training and testing, and verb phrase training and testing are identical to those in Experiment 1. The instructions for adjective training and testing, and the critical noun phrase testing are as below:
Adjective training phase: In each trial, you will see four pictures. In each picture, the objects are the same color. You will hear a word to describe the color. Repeat what you hear, then press the spacebar to continue.

Instruction for each trial: Repeat what you hear, then press the spacebar to continue.

Adjective testing phase: In each trial, you will see four pictures as before. Say the name of the color, and press the spacebar, then you will hear the correct answer.

Instruction for each trial: Describe the color the objects share, then press the spacebar to continue.

Noun phrase testing phase: This is your last test. In each trial, you will see a picture of an object. Describe the full picture. You will need to use two words. You have not heard descriptions like this before, so just do your best. When you are finished, press the spacebar to continue.

Instruction for each trial: Describe the full picture, then press spacebar to continue.

C.1.3 Instructions for Experiment 3

The instructions for adposition condition in Experiment 3 are identical to those in Experiment 1. The instructions for color adjective condition in Experiment 3 are identical to those in Experiment 2. The welcome text and the instructions for noun training and testing, verb training and testing, verb phrase training, verb phrase matching, verb phrase production testing, and noun phrase testing in texture adjective condition and active adjective condition in Experiment 3 are identical to those in Experiment 2. The instructions for adjective training and testing in texture adjective condition and active adjective condition are the same, which are shown below:

Adjective training phase: In each of the following trials, you will see four pictures. In each picture, the objects share the same property. You will hear a word to describe this property that they all share. Repeat what you hear, then press the spacebar to continue.
Instruction for each trial: Repeat what you hear, then press the spacebar to continue.

Adjective testing phase: In each of the following trials, you will see four pictures as before. Say the name of the property, and press the spacebar, then you will hear the correct answer. Now press the spacebar to start.

   Instruction for each trial: Describe the property the objects share, then press the spacebar to continue.

C.2 Instructions for experiments in Part II

C.2.1 Instructions for Experiments 1A and 1B

Welcome text: Welcome! You are going to learn a new language named Bahareti. It is spoken by roughly 50,000 people in the Southeast Asia. It has some different features from English. Your participation will help us to understand how English speakers learn and use this language. Simply try your best to learn this language. Instructions will be provided at the beginning of each section. Please read and follow the instructions carefully. Now please press the spacebar to start.

Noun training phase: In the following trials, you will see some objects on the screen along with the Bahareti words for these objects. Try to learn these words. Now press the spacebar to start.

   Instruction for each trial: Press the spacebar to continue.

Noun testing phase: In the following trials, you will be tested on the words for the objects. Now press the spacebar to start.

   Instruction for each trial: Click on the word for the object.

Numeral training phase: In each of the following trials, you will see a certain number of objects on the screen. You will also see a word to express the corresponding number in Bahareti. Try to learn the number words. Now press the spacebar to start.

   Instruction for each trial: Press the spacebar to continue.
**Numeral testing phase:** In the following trials, you will be tested on the Bahareti number words you just learned. Now press the spacebar to start.

*Instruction for each trial:* Click on the word that matches the number of the objects.

**Classifier training phase:** Now you are going to learn another set of Bahareti words. Unlike English, Bahareti has special words that you must use when counting. The special words for different objects might be different. In each of the following trials, you will see four pictures. The objects in these pictures all use the same special word for counting. You will see the special word for these objects. Try to learn these special Bahareti words. Now press the spacebar to start.

*Instruction for each trial:* Press the spacebar to continue.

**Classifier testing phase:** In the following trials, you will be tested on the Bahareti words that are used when counting. Now press the spacebar to start.

*Instruction for each trial:* Click on the word that is appropriate for counting these objects.

**Mixed testing phase:** In the following trials, you will be tested on all of the words you have learned. Now press the spacebar to start.

*Instruction for noun testing trials:* Click on the word for the object.

*Instruction for numeral testing trials:* Click on the word that matches the number of the objects.

*Instruction for classifier testing trials:* Click on the word that is appropriate for counting these objects.

**Critical phrase testing phase:** In the following trials, you will see one picture and two phrases. Note that Bahareti speakers need to use all three of the types of words you have learned to describe each of these pictures. Click on the phrase that you think a Bahareti speaker would use to describe the picture. Now press the spacebar to start.

*Instruction for each trial:* Click on the phrase that you think a Bahareti speaker would use to describe the picture.
C.2.2 Instructions for Experiment 2

The welcome text and the instructions at the beginning of each phase are identical to those in Experiments 1A and 1B. The instructions for trials in different phases are as below:

*Instruction for noun training trials:* Repeat the word you hear, then press the spacebar to continue.

*Instruction for noun testing trials:* Say the label for the object, then press the spacebar to continue.

*Instruction for numeral training trials:* Repeat the word you hear, then press the spacebar to continue.

*Instruction for numeral testing trials:* Say the number of the objects, then press the spacebar to continue.

*Instruction for classifier training trials:* Repeat the word you hear, then press the spacebar to continue.

*Instruction for classifier testing trials:* Say the special word used when counting these objects, then press the spacebar to continue.

*Instruction for phrase testing trials:* Repeat the noun first and then complete the phrase to describe the picture. Then press the spacebar to continue.

C.2.3 Instructions for Experiment 3

The welcome text is the same as that in Experiments 1A, 1B, and 2.

**Noun training phase:** In the following trials, you will see some objects on the screen along with the Bahareti labels for these objects. Try to learn these object labels. Now press the spacebar to start.

*Instruction for each trial:* Try to learn the label for the object, then press the spacebar to continue.

**Noun testing phase:** In the following trials, you will be tested on the labels for the objects. Now press the spacebar to start.

*Instruction for each trial:* Say the object label in Bahareti, then press the spacebar to continue.
Numeral and classifier training phase: In the following trials, you will learn two types of words that are needed to describe the quantity of objects in Bahareti. Try to learn these words. Now press the spacebar to start.

*Instruction for numeral training trials:* Objects in the two pictures **differ in number** (but have the same measure unit). Press the spacebar to continue.

*Instruction for classifier training trials:* Objects in the two pictures **differ in measure unit** (but have the same number word). Press the spacebar to continue.

Numeral and classifier matching phase: In the following trials, you will see a word and two pictures. Click on the picture that matches the word. Now press the spacebar to start.

*Instruction for numeral matching trials:* Click on the correct picture that matches the number word.

*Instruction for numeral matching trials:* Click on the correct picture that matches the measure unit.

Numeral and classifier testing phase: In the following trials, you will be tested on the two types of words you just learned. Now press the spacebar to start.

*Instruction for numeral testing trials:* Say the number of the highlighted objects in Bahareti, then press the spacebar to continue.

*Instruction for classifier testing trials:* Say the measure unit of the highlighted objects in Bahareti, then press the spacebar to continue.

Mixed testing phase: In the following trials, you will be tested on all of the words you have learned. Now press the spacebar to start.

*Instruction for noun testing trials:* Say the object label in Bahareti, then press the spacebar to continue.

*Instruction for numeral testing trials:* Say the number of the highlighted objects in Bahareti, then press the spacebar to continue.

*Instruction for classifier testing trials:* Say the measure unit of the highlighted objects in Bahareti, then press the spacebar to continue.

Critical phrase testing phase: Well done! This is your last test. In each of the following trials, you will see one picture. Note that Bahareti speakers need to use
all three of the types of words you have learned to describe the picture. Object
labels in Bahareti always come first. Now press the spacebar to continue.

Here is a hint about how to do this: first you say the object label and then the
other two words. Just guess what you think a native Bahareti speaker would do!
Now press the spacebar to start.

*Instruction for each trial:* The object label comes first. Then the other two words.
Then press the spacebar to continue.
Bibliography


Allassonnière-Tang, Marc, Olof Lundgren, Maja Robbers, Sandra Cronhamn, Filip Larsson, One-Soon Her, Harald Hammarström & Gerd Carling. 2021. Expansion by migration and diffusion by contact is a source to the global diversity of linguistic nominal categorization systems. *Humanities and Social Sciences Communications* 8(1). 1–6.


Christiansen, Morten H & Joseph T Devlin. 1997. Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations.


Culbertson, Jennifer, Marieke Schouwstra & Simon Kirby. 2020b. From the world to word order: deriving biases in noun phrase order from statistical properties of the world. *Language* 96(3). 696–717.


Dryer, Matthew S. 2013g. Relationship between the order of object and verb and the order of adjective and noun (v2020.3). In Matthew S. Dryer & Martin Haspelmath (eds.), The world atlas of language structures online, Zenodo.


Durvasula, Karthik & Adam Liter. 2020. There is a simplicity bias when generalising from ambiguous data. *Phonology* 37(2). 177–213.


191


Jenks, Peter. 2010. Evidence for the syntactic diversity of numeral classifiers. In LSA Annual Meeting extended abstracts, vol. 1, doi: https://doi.org/10.3765/exabs.v0i0.494.


LaPolla, Randy J. 2002. Problems of methodology and explanation in word order universals research. In Wuyun Pan (ed.), *Dongfang yuyan yu wenhua (Languages and cultures of the east)*, 204–237. Shanghai: Dongfang Chuban Zhongxin.


Little, Carol-Rose, Mary Moroney & Justin Royer. 2022. Classifiers can be for numerals or nouns: Two strategies for numeral modification. *Glossa: a journal of general linguistics* 7(1).


Martin, Alexander, David Adger, Klaus Abels, Patrick Kanampiu & Jennifer Culbertson. 2024. Not all cross-linguistic generalizations are driven by universal cognitive biases... but some are. Psychological Science 35(3). 304–311.


Smith, Kenny, Amy Perfors, Olga Fehér, Anna Samara, Kate Swoboda & Elizabeth Wonnacott. 2017. Language learning, language use and the evolution of linguistic variation. Philosophical Transactions of the Royal Society B: Biological Sciences 372(1711). 20160051.


