THE UNIVERSITY
of EDINBURGH

Predictive structure and the learnability of inflectional paradigms: investigating whether low i-complexity benefits human learners and neural networks

Tamar Johnson

Doctor of Philosophy

School of Philosophy, Psychology and Language Sciences

University of Edinburgh

May 2021

Predictive structure and the learnability of inflectional paradigms: investigating whether low i-complexity benefits human learners and neural networks

Tamar Johnson

ABSTRACT

Research on cross-linguistic differences in morphological paradigms reveals a wide range of variation on many dimensions, including the number of categories expressed, the number of unique forms, and the number of inflectional classes. This typological variation is surprising within the approach that languages evolve to maximise learnability (e.g., Christiansen and Chater 2008; Deacon 1997; Kirby 2002). Ackerman and Malouf (2013) argue that there is one dimension on which languages do not differ widely: in predictive structure. Predictive structure in a paradigm describes the extent to which forms predict each other, sometimes called i-complexity. Ackerman and Malouf (2013) show that although languages differ according to surface paradigm complexity measures, called e-complexity, they tend to have low i-complexity.

While it has been suggested that i-complexity affects the task of producing unknown forms (the Paradigm Cell Filling Problem, Ackerman, James P. Blevins, et al. 2009; Ackerman and Malouf 2015), its effect on the learnability of morphological paradigms has not been tested. In a series of artificial language learning tasks both with human learners and LSTM neural networks, I evaluate the hypothesis that learners are sensitive to i-complexity by testing how well paradigms which differ on this dimension are learned.

In Part 1, I test whether learners are sensitive to i-complexity when learning inflected forms

in a miniature language. In Part 2, I compare the effect of i-complexity on learning with that of e-complexity and assess the relationship between these two measures, using randomly constructed paradigms. In Part 3, I test the effect of i-complexity on learning and generalisation tasks, manipulating the presence of extra-morphological cues for class membership.

Results show weak evidence for an effect of i-complexity on learning, with evidence for greater effects of e-complexity in both human and neural network learners. A strong negative correlation was found between i-complexity and e-complexity, suggesting that paradigms with higher surface paradigm complexity tend to have more predictive structure, as measured by i-complexity. There is no evidence for an interaction between i-complexity and extra-morphological cues on learning and generalisation. This suggests that semantic or phonological cues for class membership, which are common in natural languages, do not enhance the effect of i-complexity on learning and generalisation. Finally, i-complexity was found to affect generalisation in both human and neural network learners, suggesting that i-complexity could, in principle, shape languages through the process of generalisation to unknown forms.

I discuss the difference in the effects of i-complexity on learning and generalisation, the similarities between the effect of i-complexity in human learners and neural networks, and cases the two types of learner differed. Finally, I discuss the role that i-complexity is likely to have in language change based on the results.

Predictive structure and the learnability of inflectional paradigms: investigating whether low i-complexity benefits human learners and neural networks

Tamar Johnson

GENERAL AUDIENCE ABSTRACT

Languages are thought to adapt to their learners' cognitive abilities and to simplify over time. Under this perspective, it is surprising that some languages seem very difficult to acquire. For instance, nouns in many languages are modified to express additional information such as the number of objects, the noun's role in the reported action and its relationship to other nouns in the sentence. Moreover, there could be several different ways to mark each type of information, depending on the noun. In Greek, for example, nouns are modified for number (singular or plural) but are also modified for whether the noun is the direct object of the action (e.g., in the sentence "the girl eats the apple", *the apple* is the direct object). A direct object noun is modified with one of the endings *-o, -on, -os* or with no additional ending, depending on the noun itself.

In this thesis I explore whether language learners benefit from cases where the way words in the language are modified for marking specific information (for example a plural form of a direct object noun) can be predicted based on how they are modified for marking other types of information (the singular form of a direct object noun, for instance). Taking an example from Greek, a direct object noun that is modified with the ending *-on* in singular, is always modified with the ending *-us* as a direct object noun in plural, and not with any other endings conveying the same information (*-es, -is, -a* or *-i*).

A language in which words can better predict how other words are modified to mark different information is seen as less complex due to its predictive structure. Here, I test whether languages featured with predictive structure are more easily learned than languages with no such structure, all other things being equal. If language learners do use these predictive relationships between words when acquiring a new language, even languages that seem very complex at first can still be easily learned.

I explore this suggestion by training and testing learners on miniature artificial languages that I designed. Half of the languages learners are trained on are featured with predictive structure, and the other half do not. I then test if languages with the predictive structure were indeed easier to learn. The learners in my experiments were human participants completing the task on an online platform. In addition to these human learners, I also trained and tested computational models of learning (neural networks) on the same languages.

My results show that the predictive structure of languages does not strongly affect the difficulty of learning the language. In most of the tasks I used here, the computational learning models and human learners displayed similar weak effects of the predictive structure on their learning. I conclude that probably other features of the language affect its learnability more than its predictive structure.

Acknowledgments

This thesis would not exist without the guidance from my supergroup of supervisors Kenny Smith, Jennifer Culbertson and Hugh Rabagliati. I feel privileged to have had the opportunity to learn from these brilliant people. The qualities of the academic environment created by them - those of mutual respect, support, curiosity, and growth - are ones that I will always seek out in my professional career. Kenny has the unique ability to be attuned to the smallest details in research projects as he is to those who work with him. He was attentive to the challenges I faced over the course of this PhD and would offer words of encouragement to help me see them through. Jenny always had high expectations of my work and challenged me to become a better scholar. Being a talented writer, I see her as a role model for how to communicate my ideas clearly and coherently. Hugh would get me to zoom out from the trivialities of research to envision my future career in academia. He was also the person I could go to with statistics-related quandaries, which he helped me solve time and again.

I am thankful for the thought-provoking weekly meetings with the members of the Evolution of Linguistic Complexity research group. It was a great pleasure to work alongside each of them and to get their valuable advice: Stella Frank, Carmen Saldaña, Jia Loy, Helen Sims-Williams, Clem Ashton, Rachel Kindellan and Juliet Dunstone. I can only hope I will be lucky enough to enjoy such friendly and fruitful collaborators in the future.

The Centre for Language Evolution and its bright members gave me a fertile ground to explore and develop professionally within a unique interdisciplinary community.

I also want to thank my peers that made my time in Edinburgh more pleasant and memorable: Fausto Carcassi, Jon Carr, Henry Coxe-Conklin, Annie Holtz, Andres Karjus, Fiona Kirton, Mora Maldonado, Alex Martin, Marc Meisezahl, Danielle Naegeli, Jonas Nölle, Asha

Sato, Svenja Wagner, Fang Wang, and Marieke Woensdregt. I feel lucky to have spent the past few years with such exceptional individuals. I hope to see you all soon for a drink at Dagda's.

Nina, Julian and Olivia, you were the best flatmates. You helped me find my place in Edinburgh and feel at home. I am grateful for my family, their support over these years, and especially for my mom and dad, Miri and Neil Johnson, who taught me to dream and follow my curiosity and heart in every step I take.

This thesis was completed during a global pandemic and it would not have been possible without the endless patience and encouragement from my partner, Amit, who always sought to assist me in pursuing my dreams and to give me the best conditions to achieve them.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

Languages differ widely in their morphological systems, including in their inflectional paradigms; some languages do not use morphology to mark grammatical information at all (e.g., Mandarin) whereas others make use of inflectional morphology to mark dozens of grammatical functions (e.g., Arabic).

Linguistic variation and typological patterns are often explained in terms of the learnability of the system. This link is almost intuitive; as put by Elsner et al. (2019): "all natural languages must be learned, and 'unlearnable' linguistic systems cannot survive. Therefore, the learning mechanism provides constraints on what sorts of languages can exist in the world." An evolutionary approach to language change redefines this link, through the general hypothesis that languages evolve to maximize their learnability (e.g., Christiansen and Chater 2008; Deacon 1997; Kirby 2002). As natural languages are passed from person to person (a process referred to as cultural transmission, e.g., Kirby, Cornish, et al. (2008)) languages become more and more learnable. Under this approach, high linguistic variation is surprising, especially the existence of large morphological systems such as in Arabic; we would expect to see shared features across languages that make them more learnable. For instance, dominant word orders in natural languages were shown to be more learnable (e.g., Culbertson, Smolensky, et al. 2012).

An inductive bias for simplicity is often proposed as shaping languages (Chater, Clark, et al. 2015; Chater and Vitányi 2003; Culbertson and Kirby 2016; Feldman 2003; Kirby,

Tamariz, et al. 2015; Pothos and Chater 2002). According to this view, simplicity is a general principal guiding learning across cognitive domains: learners are biased towards inferring simpler hypotheses to explain observed data. In the linguistic domain, this principal has been described as a preference for languages whose grammars can be expressed more compactly (Culbertson and Kirby 2016). The notion of linguistic complexity is not easily definable or measurable. Specifically, in morphology, there is no widely accepted measure of morphological complexity (e.g., Baerman et al. 2015; Sampson et al. 2009). Different definitions of complexity will make different predictions about what is complex and therefore what should be hard to learn. Hence, we need to look at measures of complexity.

1.1 Measures of morphological complexity

In this section I review two groups of measures of morphological complexity suggested in the literature. The aim of this overview is to explore the interaction between complexity and the learnability of morphological systems.

1.1.1 General measures of morphological complexity

Systems with inflectional morphology are thought to be more complex than ones without inflectional morphology (McWhorter 2001), since inflection systems usually create phenomena that add processing load, such as morphophonemics, supplition, declensions with arbitrary allomorphy and agreements.

Counting the grammatical categories that a lexeme in the language can be marked with often serves as a proxy for the complexity of the system (Bickel and Nichols 2013; Shosted 2006; Xanthos and Gillis 2010); the more information is marked using inflection, the more complex

the system is. Bickel and Nichols (2013) quantify the degree to which verbs can be marked by inflectional categories (e.g., tense, voice and agreement). Their measure, categories per word (cpw), is based on maximally inflected verb forms, and give the number of inflectional categories to the verb.

Other, corpus-based measures, were suggested to compute a language's morphological complexity directly based on texts rather than relying on experts' judgments. Type-token ratio (TTR) (Kettunen 2014; Malvern et al. 2004), reflects the richness of the morphological paradigm, by calculating the ratio of the number of distinct words (types) to the total number of words (tokens); the more information is marked morphologically in the language, the more different types are expected to be in the text.[1]

Juola (1998) and Juola (2008), suggests a measure for morphological complexity that uses the notion of Kolmogorov complexity (Chaitin 1988; Kolmogorov 1968), according to which the complexity of an object is the length of its shortest description (i.e., a compression of the language). To measure morphological complexity this method computes the ratio of the length of a compressed language, to the length of a compressed deformed version of the language; in the deformed version, all words are replaced with integers so that each occurrence of a word is replaced with the same integer. This ratio, then, measures how much information is stored within words in the language (via morphology), as in the original text, versus the information that is conveyed using separate words in the lexicon (the only information in the deformed version). Higher ratios would represent languages with richer and more complex morphology.

Another method using the notion of Kolmogorov complexity was suggested by Goldsmith (2001). Using *minimum description length* (MDL, Rissanen 1984) to approximate Kol-

[1] A moving-average type/toke ratio (MATTR) is used to control for corpus length (Covington and McFall 2010). MATTR is computed as the mean TTR's of moving windows in the text (i.e., for a window size of 500 words, for instance, it calculates the TTR of words 1-500, 2-501, 3-502 etc.).

mogorov complexity, Goldsmith (2001) proposes a method that morphologically analyses text using automated techniques to form a compressed representation of the language. The language representations in this case are lexica consisting of the stems in the language, and the distribution of affixes each stem can take (called "signatures"). A metric for morphological complexity based on this method was proposed by Bane (2008) and can be measured as in 1.1.

$$\frac{DL(Affixes) + DL(signatures)}{DL(Affixes) + DL(signatures) + DL(stems)} \tag{1.1}$$

Where $DL(x)$ is the description length of x. If the total complexity of the lexicon is distributed between its stems, affixes and signatures, its morphological complexity would be the complexity of its affixes and their distribution (signatures); for languages with fewer inflections, most of the information (complexity) would be encoded in different stems rather than in its affixes, leading to low values of this measure.

All of the measures presented above count or compute morphological complexity in the spirit of McWhorter (2001) intuition, namely that a linguistic system is more morphologically complex if it makes more extensive use of inflectional morphology. Indeed, these measures were found to correlate with each other when applying them to number of different languages and texts (Bane 2008; Bentz et al. 2016; Juola 2008; Kettunen 2014). Measures following this notion of complexity display high variation in morphological systems in natural languages (Ackerman and Malouf 2013). Therefore these measures offer little explanation, in terms of learnability, for the observed typological variation.

Learnability, however, is not the only factor shaping linguistic systems: languages are used for communication, and linguistic systems have been claimed to reflect a trade-off between inductive biases (e.g., for simplicity) and pressure from communication (e.g., minimizing ambiguity, Kemp and Regier 2012). This trade-off has been shown in a variety of linguistic

domains, where natural languages show a near optimal balance between these two pressures (e.g., Regier et al. 2015; Xu et al. 2016; Zaslavsky et al. 2020). Evidence for this trade-off has also been found in experimental studies manipulating the relative importance of learning and communication (e.g. Kirby, Tamariz, et al. 2015; Motamedi et al. 2019; Silvey et al. 2015).

While some of the variation in natural languages can be theoretically explained by the simplicity-informativeness trade-off (number of inflectional categories as measured by Bickel and Nichols (2013) for example, adds information to the verb), some factors that show high variation in morphological systems are left unexplained. For instance, number of inflectional classes vary significantly across languages (Greville G. Corbett 2005). The complexity that originates from high number of inflection classes does not add to the informativity of the system (Baerman et al. 2010).[2].

Going back to the discussion on morphological complexity and its reflection in typological variation, a second notion of complexity suggested in the literature concerns the predictive structure of morphological paradigms.

1.1.2 Predictive structure as a measure of morphological complexity

A separate approach for defining morphological complexity is the Word and Paradigm family of theories. Under this approach, the word, rather than stems and affixes, is the basic unit of morphological structure, and the focus of analysis is on the relationships between forms of the lexeme (Ackerman, James P. Blevins, et al. 2009; James P. Blevins 2006; G. T. Stump

[2]However, a number of authors have suggested that gender systems and inflection classes facilitate communication through assisting the hearer in comprehension of the noun that follow the gender-matched determiner (e.g., Dye et al. 2017)

2001). The notion of morphological complexity reflected in these measures is the difficulty with which language users can predict an unobserved inflected form of a lexeme, based on knowledge of another inflected form of the lexeme. This is the Paradigm Cell Filling Problem (PCFP) (Ackerman, James P. Blevins, et al. 2009). To illustrate, take for example the nominal plural paradigm in Modern Greek (Table 1.1), that expresses four cases, with a series of different morphemes that depend on the inflectional class of the word. Morphological complexity in the sense of solving the PCFP, can be exemplifies by the difficulty in guessing how a lexeme is inflected for one grammatical category (a combination of number and case), if the inflected form in another category is known. For instance, the uncertainty in guessing the form of a lexeme in Plural.Accusative, if it is known that the lexeme takes the suffix *-es* in Plural.Nominative is low (it should take *-es*), whereas the uncertainty based on knowledge that the lexeme takes *-on* in Plural.Genitive is high.

	Plural			
class	**Nom**	**Gen**	**Acc**	**Voc**
1	-i	-on	-us	-i
2	-es	-on	-es	-es
3	-es	-on	-es	-es
4	-is	-on	-is	-is
5	-a	-on	-a	-a
6	-a	-on	-a	-a
7	-i	-on	-i	-i
8	-a	-on	-a	-a

Table 1.1: Modern Greek plural nominal inflectional classes (Ackerman and Malouf 2013 based on Ralli (2002)). Columns give the inflectional endings for nouns in different grammatical roles (nominative, genitive, accusative and vocative), for plural; rows show the 8 inflectional classes in Modern Greek.

Measures of morphological complexity reflecting this view, include set-theoretic (G. Stump and R. A. Finkel 2013) and information-theoretic (Ackerman and Malouf 2013; Bonami and Beniamine 2016; Cotterell, Kirov, Hulden, et al. 2019; Sims and Parker 2016; G. Stump and R. A. Finkel 2013) measurements to examine the predictability of forms in inflectional

paradigms.

G. Stump and R. A. Finkel (2013) define the term *principal parts* as the smallest set of inflected forms that needs to be known in order to infer correctly all other inflected forms for the same lexeme. In other words, principal parts are the smallest number of forms that allow the language user to infer the lexeme's inflection class; for languages with no inflection classes, where all lexemes of the same part of speech inflect alike, there is no use of principal parts as the entire paradigm can be deduced even without knowledge of any form of the lexeme (only to deduce its stem or root).

Ackerman and Malouf (2013) proposed a measure with the same objective, reflecting the difficulty of solving the PCFP, using the notion of entropy (Shannon 1963) from information theory, and name it *integrative complexity (i-complexity)*. In general, entropy reflects the amount of uncertainty in predicting a value for an object X, based on the probabilities of x, the values X can take . Defined as in 1.2.

$$H(X) = -\sum_{x \in X} P(x) log_2 P(x) \tag{1.2}$$

Take for instance X as the part of speech of a word in a text in English and x as the possible values it can take (noun, pronoun, verb, adjective etc.). If *noun* is the most common part of speech in the text with the rest of options appearing in low probability, the uncertainty or entropy of X is low. However, if all possible parts of speech have the same probability (i.e., appearing in the text in similar proportions), the uncertainty and entropy of X is high.

Specifically, Ackerman and Malouf (2013) use the notion of conditional entropy of forms in the paradigm for measuring morphological complexity. Conditional entropy reflects the uncertainty in predicting the value of an object Y, given the value of another object X, and can be defined as in 1.3.

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} P(y,x) log_2 P(y|x) = -\sum_{x \in X} \sum_{y \in Y} P(y,x) log_2 \frac{P(y,x)}{p(x)} \tag{1.3}$$

Where P(y,x) is the joint probability of both X and Y to be realized with the values x and y respectively. As in the example with English parts of speech, H(Y|X) would be the entropy of the part of speech of word Y given the part of speech of the preceding word, X. P(y,x) would be the joint probability over all values of X and Y to follow each other in the text.

To measure morphological complexity of an inflectional paradigm, Ackerman and Malouf (2013) propose using average conditional entropy over all inflection categories in the paradigm. Conditional entropy reflects the uncertainty in predicting how a lexeme is realized in inflection category Y, given the realization of the lexeme in the inflection category X; here P(y,x) indicates the joint probability of two inflection categories in the paradigm being realized as forms y and x, respectively. For illustration, looking at the nominal inflection paradigm in modern Greek (Table 1.1), the uncertainty in guessing the realization of a lexeme in plural.Accusative is high (it can take either one of the affixes *-us, -es, -is, -a* or *-i*), however, if the realization of the lexeme in Plural.Nominative is known (e.g., it takes the suffix *-is*), then the uncertainty is much lower (it takes *-is* in Plural.Accusative as well).

Average conditional entropy is the mean conditional entropy over all pairs of inflection categories in the paradigm, as in 1.4.

$$\frac{\sum_{Y \in G} \sum_{Y \in G, X \in G, X \neq Y} H(X|Y)}{N_G(N_G - 1)} \tag{1.4}$$

Where G is the set of inflectional categories in the paradigm and N_G is their total number.

I-complexity has been proposed as a measure of morphological learnability within the Word-and-Paradigm approach. However, some criticize the use of average conditional entropy as

a formal model of i-complexity, as proposed by Ackerman and Malouf (2013). The criticism concerns both *averaging* the conditional entropy of the paradigm and segmenting the inflected forms into stems and affixes (or exponents) for applying analogy over affixes only (as exemplified in Table 1.1). In this context, sequence-to-sequence computational models were proposed to simulate the difficulty of predicting inflections of a lexeme (i.e., estimating i-complexity) as a more fine-grained approximation than average conditional entropy. The task of *morphological reinflection* (i.e., the conversion of one inflected form to another) is commonly used in these models as a computational formalization of morphological predictability (e.g., Cotterell, Kirov, Hulden, et al. 2019; Cotterell, Kirov, Sylak-Glassman, et al. 2016; Malouf 2017, see Elsner et al. (2019) for discussion).

While acknowledging this criticism over the use of average conditional entropy as a measure of i-complexity, the emphasis in this thesis is on testing whether inter-predictablility of forms in an inflection paradigm, as a main aspect of i-complexity, affects the learnability of the paradigm.

Ackerman and Malouf (2013) distinguish i-complexity from measures that reflect the amount of information that is conveyed using morphology and the strategies employed to encode this information, including allomorphy over inflection classes. They refer to the second type of complexity as *enumerative complexity (e-complexity)*. Over different languages it has been shown that i-complexity is consistently low, while e-complexity varies widely (Ackerman and Malouf 2013; Bonami and Beniamine 2016; Cotterell, Kirov, Hulden, et al. 2019; G. Stump and R. A. Finkel 2013; Wilmoth and Mansfield 2021). Ackerman and Malouf (2013) calculate i-complexity for inflectional paradigms in a set of 10 geographically and genetically varying languages. The languages' e-complexity varied widely, while their i- complexity values were under 1 bit across the board. A simulation analysis performed on one of the languages exhibiting high e-complexity (Chiquihuitlàn Mazatec) showed that the i- complexity of the

actual paradigm was lower than the i-complexity values for random permutations of that paradigm. Based on these findings, Ackerman and Malouf postulate the *Low Conditional Entropy Conjecture*; they suggest that the inflectional paradigms of natural languages are organized in such a way as to minimize their i-complexity. The Low Conditional Entropy Conjecture helps reconcile the surprising variation found in languages with respect to their e-complexity; inflectional systems of natural languages may be organized according to their i-complexity rather than other parameters.

Does this typological pattern reflect an inductive bias, i.e., are there advantages in learning for low i-complexity languages? The central question in this thesis is whether low i-complexity facilitate learning of inflectional paradigms.

1.2 Previous evidence for the effects of i-complexity on learning

The high variation in inflectional paradigms of natural languages is intuitively in conflict with the general hypothesis that languages evolve to maximize their learnability (e.g., Christiansen and Chater 2008; Culbertson and Kirby 2016; Deacon 1997; Kirby 2002; Kirby, Cornish, et al. 2008. However, the Low Conditional Entropy Conjecture (Ackerman and Malouf 2013) might suggest that inflectional paradigms are organized according to their i-complexity in order to accommodate for the users' bias for simplicity.

A separate line of investigation has found that information-theoretic measurements of inflectional paradigms predict speakers' response times in lexical decision tasks (del Prado Martın et al. 2004; Milin, Đurđević, et al. 2009), suggesting a link between measures of this notion and language processing.

Support for the hypothesis that low i-complexity facilitates learning can be inferred from the role of analogy in language learning. Recall that i-complexity represents the extent to which inflectional forms in a paradigm can predict one another. One way for this to happen is through analogy; if a word behaves like another in one inflectional category, then by analogy it will behave like that word in another inflectional category (Ackerman and Malouf 2013; James P. Blevins 2006; James P Blevins et al. 2016; Malouf 2017; G. T. Stump 2001). Exemplar-based models of classification (e.g., Medin and Schaffer 1978; Nosofsky 1988; E. E. Smith and Medin 2013 and more recently Ambridge 2020) suggest that human learners store exemplars in memory and categorization decisions are made by relying on similarities between target and stored items. Indeed, a number of previous studies suggest that adults and children will choose inflections for novel words based on their phonological similarity to familiar words (e.g., Ambridge 2010; Milin, Keuleers, et al. 2011). Although these accounts mostly look at similarities in stems rather than similarities in inflectional behaviour, as represented by i-complexity, they are compatible with the idea that learners build relations between forms in part by analogy.

Furthermore, there is evidence that i-complexity affects the generalization of the paradigm to novel words. Seyfarth et al. (2014) tested Ackerman, James P. Blevins, et al. (2009) hypothesis that i-complexity has an effect on the ability of human learners to solve the Paradigm Cell Filling Problem. They compared the ability of human learners to predict novel inflected forms in low vs. high i-complexity input. They trained participants on an artificially constructed nominal inflectional paradigm in which nouns were marked for three grammatical numbers (singular, dual and plural) according to one of two noun classes (Table 3.2(a)). In the test phase, they asked participants to generate inflected forms for a novel lexeme given that lexemes' inflected form in another grammatical number. In some trials, the required form could be predicted from the given form (predictive trials), while in

others it could not be (non- predictive trials). In Table 3.2(a) for example, being prompted with a novel singular form marked with -*yez* allows the learner to predict what form the lexeme takes in the dual (-*cav*). However, knowing the form in plural is not predictive of the form in dual. They found that participants' performance differed across predictive and non-predictive trials, showing that learners were indeed able to use the predictive structure to generate a correct novel form. In a second experiment, Seyfarth et al. (2014) tested whether predictive information facilitated generalization to novel stems in a larger paradigm (Table 3.2(b)). They found that learners made less use of predictive information in this larger paradigm: learners tended to inflect novel stems with the most frequent marker (e.g., they used the suffix -*cav* to mark dual regardless of class).

Table 1.2: Artificially constructed nominal inflection paradigms used in Seyfarth et al. (2014).

	Singular	Dual	Plural
noun class 1	-yez	-cav	-lem
noun class 2	-taf	-guk	-lem

(a) Paradigm with two noun classes (their Experiment 1).

	Singular	Dual	Plural
noun class 1	-taf	-guk	-lem
noun class 2	-yez	-cav	-lem
noun class 3	-yez	-cav	-nup

(b) Paradigm with three noun classes (their Experiment 2).

Seyfarth et al. (2014) show that learners are able to generalize inflectional paradigm based on predictive structure, suggesting that low i-complexity can be used for solving the PCFP.

I attempt to explore whether the typological patterns of i-complexity reviewed above reflect an inductive bias through testing it with two types of learners, human learners and LSTM neural networks, in learning inflectional paradigms and generalizing them to novel forms. Furthermore, to evaluate the effect of i-complexity on learning inflectional paradigms, I compare its effect with the effect of e-complexity, using a measure I propose here. In the

next section I motivate each of these main exploratory themes.

1.3 Three main exploratory themes

1.3.1 The task of learning inflectional paradigms

The Paradigm Cell Filling Problem (Ackerman, James P. Blevins, et al. 2009) reflects the task that the language user faces when having to produce a completely novel form based on other forms they have encountered. However, generalizing to completely novel forms is an extreme case of a much more general problem that language learners face. In addition to generalizing to completely novel forms, learners must generate (or retrieve) forms which may have been encountered but have not yet been robustly acquired. My focus in this thesis is on the more general task of retrieving low frequency forms; testing whether knowledge of other inflected forms of the same lexeme assists in this task, when the inflectional paradigm has predictive structure.

The general task language users and learners face can be used to test whether i-complexity has a role in language change through the learners' inductive bias. Based on evidence that low i-complexity facilitates solving the Paradigm Cell Filling Problem (Seyfarth et al. 2014), i.e., using familiar forms to predict new forms, I hypothesize here that it should, in principle, facilitate learning forms under low exposure as well; learners can use the same strategy they use when generalizing to completely novel stems to help generate (or recall) low frequency forms that are not fully memorized.

For testing i-complexity as reflecting users' inductive bias, I focus on estimating how easily the paradigm is learned, i.e., measuring the amount of trials or the overall accuracy in retrieving trained forms. In addition, I replicate Seyfarth et al. (2014) results on the effect

of i-complexity in generalization with human learners and show similar results with neural networks.

1.3.2 Using neural networks as psycholinguistic subjects

Throughout the thesis, I use LSTM networks as a supplement to human learners as an additional means of testing the relative impact of i-complexity on paradigm learning. This method is part of an attempt to systematically investigate the role i-complexity might have in facilitating morphological learning.

The motivation to use neural networks as computational models of language acquisition and processing dates back to Rumelhart and McClelland (1986), who attempted to present a learning model for the case of English past tense. They argued against the hypothesis that learning has to be ruled-based, presenting a neural network model whose task is to capture both regular and irregular verbs under no explicit rules of inflection. However, their model's results were heavily criticized by Pinker and Prince (1988) who pointed out theoretical and empirical failures of the model, demonstrating why it cannot be a representative model of human cognition. Following the high impact criticism on Rumelhart and McClelland (1986) model, linguists were reluctant to use neural networks as models of language learning, despite the significant improvements to neural networks' processing abilities (Jeffrey L. Elman 1990; Jordan 1997) and the growing popularity they gained in cognitive science more generally (Bechtel and Abrahamsen 1991; Jeffrey L Elman et al. 1996; McCloskey 1991).

More recently, there has been renewed interest in testing whether neural networks can represent human processing and learning in psycholinguistic tasks, following the success of these models in the field of natural language processing (NLP). Specifically, recurrent neural networks with Long Short Term Memory units (LSTM), were shown to be capable of achieving

performance comparable to humans in psycholinguistics tasks (e.g., Futrell et al. 2019; Gulordava et al. 2018; Kirov and Cotterell 2018; Linzen et al. 2016; McCurdy et al. 2020). For the task of representing hierarchical information in sequence processing, Linzen et al. (2016) show that LSTM networks can in some cases predict long-distance subject-verb number agreement, in the presence of other potential agreement triggers (often called attractors) intervening between the subject and verb; Gulordava et al. (2018) show that LSTMs trained on four different languages can often accurately predict subject-verb agreement even when they are not trained specifically on that task; Futrell et al. (2019) show that surprisal scores of LSTMs (a measure of processing cost) paralleled preferences of human participants on grammatical judgments task differentiating word-order alternations.

Kirov and Cotterell (2018) used recurrent neural networks with LSTM units in an encoder-decoder architecture to test their performance in English past tense task. Their model show near-ceiling performance in generalizing the regular past-tense suffix /-(e)d/ to held-out test data. The errors produced by the model were similar in pattern to those made by human language learners; the model, like humans, tend to overuse the regular past tense form. However, studies testing results produced with Kirov and Cotterell (2018) model more closely (Corkery et al. 2019) and studies testing the model on German number inflection (where there is no regular form) (McCurdy et al. 2020), suggest that the model produces different error patterns compared to human data.

Previous studies using recurrent neural networks with LSTM units suggest that neural networks have progressed since Rumelhart and McClelland (1986) in their ability to capture human-like behaviour in language processing. Although there are still differences in the behaviour of humans and networks in linguistic tasks, there could also be advantages to the use of LSTMs as psycholinguistic subjects.

I chose LSTMs as a reasonable default model that shows good results in a variety of language-

relevant learning problems and allows what I considered to be a natural presentation of the training data (i.e. as a sequence-to-sequence learning model). The important aspect of the use of neural networks in this thesis is using the same architecture throughout to enable the systematic examination of the effect of i-complexity in neural networks alongside human learners. In Chapter 3 I perform a sensitivity analysis to some of the default parameters and throughout I test a range of network sizes in case it affects the results.

In the thesis, I use LSTM neural networks as a convenient 'ideal learner', to test whether i-complexity can in principle influence paradigm learnability. LSTM networks have number of benefits when using them as 'subjects' in a psycholinguistic task. First, unlike human learners, neural networks display less variation across different runs of the same model. This is a very useful quality for subjects that limits the noise in the data originating from uncontrolled factors. Second, the LSTM models allow us to increase the reliability of our task; any patterns seen in both learners regarding the effect of i-complexity are less likely to be a result of uncontrolled biases, but a result of our manipulation. Finally, directly comparing performance of LSTMs and humans on a matched task opens up the possibility that, to the extent that they show similar patterns of performance, LSTMs could be used as a convenient tool to quickly generate predictions to be tested in further human experiments on paradigm learning.

1.3.3 Measuring e-complexity

Testing the effect of i-complexity on learning as well as on generalization both with human learners and neural network was done to systematically explore the role i-complexity has in shaping inflection paradigms in natural languages. Another means for doing so is by comparing the effects of i-complexity with effects of another measure of morphological complexity I

propose here. Essentially, this measure captures a notion of morphological complexity that accounts for the number of inflectional classes in a paradigm and the use of allomorphy and is measured using the notion of entropy. Crucially, this measure does not capture the predictability of forms based on other forms in the paradigm, as captured by i-complexity. I refer to this measure as e-complexity throughout, borrowing this term from Ackerman and Malouf (2013).

Ackerman and Malouf (2013) refer to e-complexity as the amount of information that is conveyed by the inflectional paradigm and the strategies employed to encode this information, including allomorphy over inflection classes. Since they do not explicitly suggest a measure for e-complexity, I propose here to adopt their *average cell entropy* as a measure for e-complexity. The average cell entropy captures the number of inflection classes and the number of different variants to mark each grammatical category (e.g., cases in the Modern Greek plural nominal inflection paradigm in 1.1).

Cell entropy is defined in (1.5) below, and is computed as the entropy of realizations for each grammatical category, X.

$$H(X) = -\sum_{x \in X} P(x) log_2 P(x) \tag{1.5}$$

It captures the difficulty with which the language user can produce the correct realization of a lexeme in a grammatical category X. Intuitively, grammatical categories that are realized with a large set of optional forms (allomorphs), or do not have a dominant/frequent variant, have higher average cell entropy, increasing uncertainty for the learner.

E-complexity is measured as the averaged cell entropy over all grammatical functions in a paradigm as in (1.6).

$$\frac{\sum_{X \in G} H(X)}{N_G} \tag{1.6}$$

Where G is the set of grammatical functions in the paradigm and N_G is their total number.

Note that the difference between i- and e-complexity rests on the extent to which they take into account the inter-predictability of forms across the paradigm. I-complexity is specifically defined to measure the degree to which one form can be guessed based on another form, in any other cell of the paradigm. In other words, it critically involves predicting the form of a lexeme in some grammatical function based on the form of that lexeme in a different grammatical function. By contrast, average cell entropy is only defined in terms of a single grammatical function, i.e., it is based on what one can predict from the form of other lexemes for that grammatical function.

The measure of e-complexity proposed here reflects a notion of morphological complexity that lies between the two approaches reviewed above (Section 1.1). On the one hand, it does not follow McWhorter (2001) notion, namely that the *amount* of information conveyed using morphology (e.g., the size of the inflectional paradigm or the maximal number of inflections a word can be marked with) reflects the complexity of the system. However, it targets the complexity that is added to the system from the existence of inflection classes and allopmorhy; one of the phenomena on which McWhorter bases his notion of complexity. Moreover, it captures an aspect of complexity that does not add to the informativity of the system (Baerman et al. 2010), such that it cannot be explained using the account of an informativity-simplicity trade-off. On the other hand, e-complexity is measured based on the notion of entropy, as does i-complexity. Yet, as noted before, e-complexity does not capture the relations between forms in the paradigm, a crucial aspect of measures in the Word and Paradigm approach (Ackerman, James P. Blevins, et al. 2009; James P. Blevins

2006; James P Blevins et al. 2016; Elsner et al. 2019; G. T. Stump 2001).

1.4 Roadmap

In Part 1, I test whether learners are sensitive to i-complexity when learning inflected forms in a miniature language. First, I give a short introduction to neural networks and present the specific architecture I am using throughout the thesis. Using this architecture, I replicate previous results with human learners showing an effect of i-complexity on generalizing inflectional paradigms to novel items. This experiment serves as a first justification for using neural networks and the specific architecture. Second, I test the effect of i-complexity on learning trained forms with neural networks and human learners on a matched artificial language learning task. To increase the likelihood of finding an effect of i-complexity, the learning task is designed to highlight the predictive structure of the paradigms; learners were trained on some of the forms in the paradigm before having to learn other forms. Finally, in order to evaluate results on i-complexity, I test the effect of e-complexity on learning inflectional paradigm and compare it to previous results with i-complexity.

To preview the main results, findings show weak evidence for an effect of i-complexity on learning, with evidence for greater effects of e-complexity in both human and neural network learners.

In Part 2, I compare the effect of i-complexity on learning with that of e-complexity and assess the relationship between these two measures, using randomly constructed paradigms. I test the effect of i- and e-complexity on learning inflection paradigms with neural networks and human learners on a learning task in which forms in the paradigm are learned in a random order. I assess the relationship between i- and e-complexity on randomly generated paradigms, i.e., when no inductive biases are in place. Moreover, Neural networks were then

trained and tested on the randomly generated paradigms to give a broader sense of how different values of the two measures affect learning. Throughout this part, other measures of morphological complexity were manipulated and their effects on learning morphological paradigms were tested as well.

Findings from Part 2 show a strong negative correlation between i-complexity and e-complexity and confirm that while neural networks are sensitive to both measures, learning is more susceptible to changes in e-complexity.

In the experiments described in Parts 1 and 2, I use artificial languages where noun class membership was not determined by the phonology or semantics of nouns. However, in many languages semantic and phonological features of nouns play a role in determining how nouns are classified. Studies show that these cues for class membership facilitate paradigm learning (reviewed in the introduction for Part 3). Therefore, in Part 3, I test the hypothesis that the effect of i-complexity on learning and generalization of inflectional paradigms interacts with the presence of phonological or semantic cues for class membership. This part extends the assessment of i-complexity and its effect on morphological learning; I test the effect of i-complexity on learning and generalisation tasks, with neural networks and human learners, manipulating the presence of extra-morphological cues for class membership.

While there is no evidence from findings from part 3 for an interaction between i-complexity and extra-morphological cues, results show that the two factors independently affect generalizing morphological paradigms.

Part I

Assessing Integrative Complexity as a predictor of morphological learning using neural networks and artificial language learning

Author Contributions

The version included here is also posted as a preprint on Arxiv. I conceived and designed the experiments and the simulations and collected the data, conducted the analysis and wrote the paper; Kenny Smith, Jennifer Culbertson and Hugh Rabagliati provided advice on the design of the experiments and data analysis, and commented on the paper.

Chapter 2

Assessing Integrative Complexity as a predictor of morphological learning using neural networks and artificial language learning

2.1 Abstract

Morphological paradigms differ widely across languages: some feature relatively few contrasts, and others, dozens. Under the view that languages are under pressure to be learnable and that the distribution of languages in the world reflects biases in language learning, this diversity is surprising – how could paradigms which apparently differ so markedly be similarly learnable? Recent work on morphological complexity has argued that even very large paradigms are designed such that they are easy to learn and use. Specifically, Ackerman and Malouf (2013) propose an information-theoretic measure, i-complexity, which captures the extent to which forms in one part of a paradigm predict forms elsewhere in the paradigm, and contrast this measure with e-complexity, which captures the number of distinctions made by the language and the different ways to mark each grammatical function. They show that languages which differ widely in e-complexity exhibit similar i-complexity; in other words,

morphological paradigms with many contrasts reduce the learnability challenge for learners by having predictive relationships between forms. Here, we test whether i-complexity predicts the learnability of inflectional paradigms using both recurrent neural networks and human participants trained on an artificial language. Furthermore, we compare the effect of i-complexity on learning with that of e-complexity. We find that in RNNs both i- and e-complexity have an effect on learning: paradigms with lower i- and e-complexity are easier to learn, although the effect of e-complexity is larger. However, for human learners, we find only weak evidence (if any) that low i-complexity paradigms are easier to learn; in contrast, low e-complexity is clearly beneficial for learning. This suggests that i-complexity might have relatively little influence on the learnability of inflectional paradigms, with other factors, such as the e-complexity having a greater effect. These results suggest that appealing to i-complexity does not fully resolve the paradox of cross-linguistic variation in morphological systems.

Keywords: morphological complexity; inflection paradigms; neural networks; LSTM; artificial language learning; entropy

2.2 Introduction

There is substantial variation in inflectional paradigms cross-linguistically. Some languages are largely devoid of inflectional morphology (e.g., Vietnamese), while others have rich inflectional systems, marking dozens of grammatical functions (e.g., Arabic). Determining the dimensions of variation in inflectional systems has been an important goal for linguistics (e.g., Bickel and Nichols 2007; Bybee 1995; Sapir 2012). However from the perspective of language learning, the existence of highly complex inflectional paradigms poses a challenge. In particular, it is often argued that languages are shaped by an inductive bias for simplicity

(Chater, Clark, et al. 2015; Chater and Vitányi 2003; Feldman 2016; Kirby, Tamariz, et al. 2015; Pothos and Chater 2002). According to this view, simplicity is a general principal guiding learning across cognitive domains: learners are biased towards inferring simpler hypotheses to explain observed data. In the linguistic domain, this principal has been described as a preference for languages whose grammars can be expressed more compactly (Culbertson and Kirby 2016). Effects of this simplicity bias on learning have been consistently reported in laboratory experiments in which learners are trained on miniature artificial languages (e.g., Canini et al. 2014; Culbertson, Smolensky, et al. 2012; Kirby, Tamariz, et al. 2015; Saffran and Thiessen 2003, a.o.). These experiments show that individual learners are less able to learn more complex linguistic patterns, infer simpler patterns whenever possible, and drive languages to become simpler over simulated generations of learners. On learnability grounds, the expectation is therefore that simple morphological paradigms should dominate cross-linguistically.[1] At first glance, this appears to be contradicted by the existence of many languages with highly complex paradigms, which have been maintained over many generations. However, in recent work, Ackerman and Malouf (2013) argue that apparently complex morphological paradigms are in fact relatively simple when complexity is measured not in terms of number of forms, but how predictable forms are from each other.

[1] Apart from learnability, other factors are also thought to shape linguistic systems: languages are used for communication, and linguistic systems have been claimed to reflect a trade-off between the bias for simplicity (i.e., minimizing the system's complexity) and pressure from communication favoring languages with higher expressivity (i.e., minimizing ambiguity, Kemp and Regier 2012). Morphological paradigms which appear highly complex could in principle reflect a balance between the communicative needs of speakers and the inductive biases of learners. However, the existence of inflectional classes–groups of lexemes that share the same set of inflectional realizations (Aronoff 1994; Greville G. Corbett 2009)–which add to the complexity of paradigms without any countervailing benefit (Baerman et al. 2010) make an analysis along these lines non-straightforward at best.

2.2.1 I-complexity and e-complexity

Ackerman and Malouf (2013) discuss two measures of complexity: enumerative complexity (*e-complexity*) and integrative complexity (*i-complexity*). The e-complexity of a language reflects the number of grammatical functions and morphosyntactic categories words in the language are marked for, the number of different forms to mark each category and their type frequencies within the morphological paradigm (Meinhardt et al. 2019). This measure of complexity is seen to express complexity according to the item and arrangement theories of morphology (Hockett 1954) that take the morpheme as the fundamental unit of analysis. Therefore, the learner's task when learning the morphological system of a language is to create an inventory of the affixes and their meanings. I-complexity is motivated by the idea that paradigms in which new forms can be easily predicted by old forms are simpler. This measure fits naturally within the word and paradigm theories of morphology in which the relationship among forms in a paradigm, and not just the forms themselves, is a crucial feature of how paradigms are represented and processed (e.g., James P. Blevins 2006; James P Blevins et al. 2016; R. Finkel and G. Stump 2007; G. T. Stump 2001). Intuitively, since exposure to lexical items is relatively sparse (i.e. a learner is unlikely to have experienced all forms of a lexeme before producing them), learners must use the forms they have heard to predict unknown forms (the Paradigm Cell Filling Problem, Ackerman, James P. Blevins, et al. (2009)). Ackerman and Malouf (2013) suggest an information-theoretic measure for I-complexity derived from Shannon entropy (Shannon 1963). It quantifies how difficult this prediction will be, namely how well one inflectional form can predict the other. This is calculated using the conditional entropy (or uncertainty) of one inflectional form Y given another X in the paradigm, as in (2.1) below:

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} P(y,x) log_2 P(y|x) = -\sum_{x \in X} \sum_{y \in Y} P(y,x) log_2 \frac{P(y,x)}{p(x)} \quad (2.1)$$

Take for example the nominal plural paradigm in Modern Greek (Table 2.1), which expresses four cases, with a series of different morphemes that depend on the inflectional class of the word. If a word takes *-i* in the nominative plural, it must be in inflectional class 1 or 7. Knowing this reduces the uncertainty of the accusative form: it must be *-us* or *-i* (not *-es*, *-is* or *-a*). By contrast, knowing that a word takes *-on* in the genitive plural does not provide any information about inflectional class, because the genitive is marked with *-on* across classes, and so it does not reduce uncertainty about the accusative form. Therefore, in Greek, H(acc.pl|nom.pl) is lower than H(acc.pl|gen.pl).

	Plural			
class	**Nom**	**Gen**	**Acc**	**Voc**
1	-i	-on	-us	-i
2	-es	-on	-es	-es
3	-es	-on	-es	-es
4	-is	-on	-is	-is
5	-a	-on	-a	-a
6	-a	-on	-a	-a
7	-i	-on	-i	-i
8	-a	-on	-a	-a

Table 2.1: Modern Greek plural nominal inflectional classes (Ackerman and Malouf 2013 based on Ralli (2002)). Columns give the inflectional endings for nouns in different grammatical roles (nominative, genitive, accusative and vocative), for plural; rows show the 8 inflectional classes in Modern Greek.

To calculate the overall i-complexity of a paradigm, one simply averages conditional entropy over all pairs of inflections as in (2.2).

$$\frac{\sum_{Y \in G} \sum_{Y \in G, X \in G, X \neq Y} H(X|Y)}{N_G(N_G - 1)} \quad (2.2)$$

Where G is the set of inflectional categories in the paradigm and N_G is their total number.

Information theory can also be used to characterise e-complexity. Specifically, we use the *cell entropy* (Ackerman and Malouf 2013), that captures the difficulty in choosing the correct inflection for a lexeme based only on the set of possible markers for that inflection in the language. Being based on entropy, this measure is more subtle than merely counting inflection classes or variants of realization in the paradigm. We therefore adopt this measure as the e-complexity of a morphological paradigm. E-complexity will be higher for paradigms with more different ways to mark each inflectional feature. It is calculated as in (2.3) below, again summing over all inflections X in a paradigm, as in (2.4).

$$H(X) = -\sum_{x \in X} P(x) log_2 P(x) \tag{2.3}$$

$$\frac{\sum_{X \in G} H(X)}{N_G} \tag{2.4}$$

Where G is the set of grammatical functions in the paradigm and N_G is their total number.

This measure captures the intuition that the Greek nominal inflection paradigm is complex because it makes a large number of distinctions and uses a variety of endings to do so. In the case of plural nouns in Modern Greek, e-complexity captures something very different from i-complexity. While the nominative forms contribute less i-complexity to the paradigm than genitives (because they are informative about other cells in the paradigm), nominative forms have higher e-complexity than genitive forms (because there are several different nominative markers but only one genitive; in fact H(gen.pl)=0).

2.2.2 Evidence for i-complexity as a predictor of learnability

Ackerman and Malouf (2013) argue that i-complexity reflects the language learner and user's task of producing an unencountered inflected forms based on other known forms in the paradigm. If solving the Paradigm Cell Filling Problem (PCFP, Ackerman, James P. Blevins, et al. 2009) reflects the learnability of the paradigm, Ackerman and Malouf (2013)'s claim regarding i-complexity generates two clear predictions. First, if two inflectional paradigms differ only in i-complexity, the paradigm with lower i-complexity should be learned more easily. Second, if paradigms with lower i-complexity are simpler to learn, they should be more common cross-linguistically (i.e., cultural transmission should lead to decreases in i-complexity which compound over generations, Kirby, Tamariz, et al. (2015)). Ackerman and Malouf (2013) provide some evidence for the second prediction: they show that, across a set of 10 geographically and genetically different languages, e-complexity varies quite widely (from 0.78 to 4.9 bits) but i-complexity is consistently low (around 0.6 bits for all languages in their sample). Moreover, they demonstrate, using a Monte-Carlo simulation on one of the languages in their sample (Chiquihuitlàn Mazatec), that the i-complexity of the verbal inflectional system of the language is minimized compared to all other permutations of the markers in the paradigm; in other words, the configuration of the paradigm is such that i-complexity is lower than we would expect if the paradigm forms were organized randomly. Both of these pieces of evidence suggest that paradigms are implicitly designed to minimise i-complexity.[2]

However, there is only limited evidence in support of the first prediction, namely, that morphological paradigms with lower i-complexity are easier to learn. Recall that i-complexity represents the extent to which one inflectional form in a paradigm can predict another. For

[2]It is worth noting that these findings have been criticized on the basis that calculating the i-complexity of languages' paradigms is highly dependent on linguists' analysis of the language and their decisions on how to describe its paradigms (e.g., Bonami and Beniamine 2016; Sims and Parker 2016).

example, if a learner can infer the inflectional class membership of a word from its realization in one inflectional category, then new inflectional forms of the word can be predicted even if they have not been previously encountered. One way for this to happen is through analogy; if a word behaves like another in one inflectional category, then by analogy it will behave like that word in another inflectional category (Ackerman and Malouf 2013; James P. Blevins 2006; James P Blevins et al. 2016; Malouf 2017; G. T. Stump 2001). Setting aside for a moment the role of i-complexity, we might ask first whether there is evidence that learners will take that analogical step. A number of previous studies suggest that adults and children will choose inflections for novel words based on their phonological similarity to familiar words (e.g., Ambridge 2010; Milin, Keuleers, et al. 2011). In addition, there is evidence from artificial language learning experiments that learners' willingness to assume that a word belongs to a particular category is influenced by perceived level of similarity among forms (e.g., Culbertson, Gagliardi, et al. 2017; Frigo and McDonald 1998; L. A. Gerken et al. 2009; Reeder et al. 2013). These results are all compatible with the idea that learners build relations between forms in part by analogy. However, they do not test the idea that predictive relationship among forms in the paradigm are critical for learning. Further, they suggest that analogy may be dependent on features of the word stems (i.e. their phonological or semantic similarity), not relations among the inflections themselves. Although Frigo and McDonald (1998) show evidence for the ability of learners to use distributional cues for novel nouns when phonological similarities are provided.

Seyfarth et al. (2014) directly test the effect of i-complexity on completing the Paradigm Cell Filling Problem – in which learners must learn a subset of forms in a paradigm, and then use those forms to predict new ones. In their Experiment 1, participants were exposed to a miniature artificial language with word forms from the nominal paradigm shown in Table 2.2a, with two inflectional classes, and three numbers. Class membership was not cued

by anything other than the pattern of inflectional endings. The paradigm's key feature is that the inflection in the dual is predicted by the word form in the singular (and vice versa), whereas the form of the plural is not predictive of any other inflectional form (since it has the same form for both inflectional classes). After learning the paradigm, participants were asked to choose the correct forms for novel stems, either in predictive trials (i.e., providing the word form in singular having been given the dual form) or in non-predictive trials (generating the singular or dual given the plural). They found that participants were more likely to generate the correct form on predictive trials than non-predictive trials. In their Experiment 2, they tested whether predictive information facilitates generalization to novel stems in a larger paradigm (Table 2.2b). Results from this experiment suggested that learners tended to inflect novel stems with the most frequent marker (e.g., they used the suffix *-cav* to mark dual regardless of class). However they did use predictive information to generalize to novel stems when the predicted suffix was low-frequency.

	Singular	Dual	Plural
noun class 1	-yez	-cav	-lem
noun class 2	-taf	-guk	-lem

(a) Paradigm with two noun classes (their Experiment 1).

	Singular	Dual	Plural
noun class 1	-taf	-guk	-lem
noun class 2	-yez	-cav	-lem
noun class 3	-yez	-cav	-nup

(b) Paradigm with three noun classes (their Experiment 2).

Table 2.2: Artificially constructed nominal inflection paradigms used in Seyfarth et al. (2014).

Further evidence comes from Malouf (2017), who trained recurrent neural networks on natural language corpora in order to test whether they could predict unobserved forms after learning partial paradigms. The networks were trained to predict phonological forms based only on a lexeme (an abstract representation of related forms without phonological informa-

tion) and a set of inflectional features (e.g. tense or number). Networks were then tested on their accuracy at producing phonological forms for lexemes with untrained inflectional features. Across a number of languages, these networks outperformed baselines which learned based on phonological information (lemmas rather than abstract lexemes) combined with specialized rules for guessing affixes. Together, these findings suggest that learners can generalize to novel inflection forms based on analogy with known forms, and further that this is facilitated by predictive links between forms in a paradigm.[3]

However, the task used in both cases serves as a relatively low bar for testing the facilitative effect of low i-complexity. For example, Seyfarth et al. (2014) compare predictive cases, where participants can generalize by analogy with learned forms, with cases where participants have no basis for generalizing (and must therefore simply guess). Therefore, it is in a sense not surprising that participants are better at determining the correct form in the former than the latter. The results of their second experiment in fact suggest that when learners have access to other cues for generalizing, e.g. marker frequency, it is less clear that learners use the kind of predictive information captured by i-complexity. More generally, both Seyfarth et al. (2014) and Malouf (2017) simulate cases in which language learners have to generalize from the partial paradigm they have learned to express an entirely new form they have never

[3]Marzi et al. (2018) give support to the hypothesis that i-complexity, rather than e-complexity, affects the learnability of the paradigm from a different angle, providing evidence that the e-complexity of the morphological system is not the main factor affecting its learnability. They propose a model for simulating learning morphological systems using a recurrent neural network trained on inflected forms and show that changes in the e-complexity of the input language does not strongly affect the learning of the morphological paradigm by the model. They trained the neural network on six languages with varying levels of e-complexity (Greek, Italian, Spanish, German, English and Arabic) and compared how successful the networks were at predicting the inflected forms across these languages. They evaluated how accurately the networks could predict incrementally presented words and found little variance in the model's results for the different languages, with a significant difference in the prediction accuracies of the model only between the most e-complex language (Modern Greek) and the least e-complex language (English). However, Marzi et al. (2018) use word prediction to test the model's learning of the forms in the paradigm without any morphological information on the inflected form given as input. Rather, the model predicts each consecutive character in the form solely based on the string of characters already presented. Therefore, it is not clear whether the results from this study provide clear evidence on the effects e-complexity has on learnability, or whether these results are instead due to the use of a task which does not include a key type of information used by learners.

been exposed to. This is an extreme form of a much more general problem that language learners face; in some cases, learners must generalize to entirely novel forms, in many others, they must generate (or retrieve) forms which may have been encountered but have not yet been robustly acquired. In principle, prediction should facilitate learning in all these cases.

2.2.3 The present study

Here we build on Seyfarth et al. (2014) by testing whether i-complexity affects how humans and neural networks learn morphological paradigms. We use artificially constructed paradigms which vary in i-complexity controlling for other factors, such as e-complexity, number of markers, and phonological similarity among stems. In a series of simulation experiments we test whether lower values of i-complexity can facilitate the learning of inflectional paradigms for recurrent neural networks (RNNs): we train Long Short Term Memory (LSTM) RNNs on two inflectional paradigms, differing only in their i-complexity values, and show that RNNs generalize more successfully to new forms in low i-complexity paradigms and that paradigms with lower i-complexity are learned faster. We then test whether low i-complexity also facilitates learning for human learners: we use the same artificial languages used for the RNNs, and train human participants on the same two inflectional paradigms, differing only in their i-complexity. In addition, we compare the effect of i-complexity on learning with that of e-complexity by training RNNs and human participants on an inflectional paradigm with high e-complexity. These experiments reveal only weak evidence of an advantage for low i-complexity languages in paradigm learning; in contrast, low e-complexity is clearly beneficial for learning. These findings suggest the possibility that although i-complexity may affect learning, it is not the central measure of morphological complexity as experienced by the language learner.

2.3 Testing the impact of i-complexity on paradigm learning in Recurrent Neural Networks

Neural network models of learning are inspired loosely by the structure of networks of neurons in the brain. Artificial neural networks consist of interconnected nodes, each with an activation value. Activation spreads from node to node via weighted connections – activation at one node will spread to other nodes with which it has positively-weighted connections, increasing their activation, and will drive down the activation of nodes with which it has negatively-weighted connections. In the types of networks we use here, nodes are arranged in layers. Nodes within one layer are connected via unidirectional weights to nodes in a subsequent layer. The first layer of nodes in the network is called the input layer and the last one is the output layer. The middle layer(s) are called hidden layers. The network weights are learned via supervised training on pairs of data points, each pair consisting of an input pattern of activation with a desired output pattern of activation. For example, the input might be the singular form of a word, and the output the corresponding plural. During training, the neural network connections are tuned with the objective of approximating the function from input patterns presented at the input layer to output patterns in the output layer. Specifically, connection weights are updated through a process called backpropagation which optimizes the weights so that when the network encounters some input pattern of activation, it produces the desired pattern of activation over the output layer. After training, the weights are fixed and the network can be tested with completely new data points or with data points similar to those presented during training (depending on the task).

Recurrent neural networks add 'short term memory' to the network, by looping back the output or hidden layer activations previously produced for earlier inputs (Jeffrey L. Elman 1990; Jeffrey L. Elman 1991; Jordan 1997). This allows networks to make predictions based

on sequences of inputs; for example, when predicting the next word in a sentence, the ability to keep previous words in memory is critical. In RNNs, the extent to which previous inputs affect the processing of the current input is also determined by weights, optimized through backpropagation.

Long Short Term Memory (LSTM) networks are an extension of recurrent neural networks introduced by Hochreiter and Schmidhuber (1997) in order to improve learning of longer temporal dependencies. Practically, LSTMs add an element of 'long term memory' to networks by allowing the network to control the influence of current and previous inputs during the process of activation propagation, using 'gates' in the networks. Like activation weights, network gates are optimized during training to determine what information is stored or passed along and therefore allowed to influence subsequent inputs. This allows LSTMs to make better use of sequential information, including learning sequential dependencies with long time lags.

A series of recent studies testing LSTM network on language tasks provide evidence that LSTM networks are capable of learning complex linguistic structure, and in some cases performance is similar to that of human participants. For example, Linzen et al. (2016) show that LSTM networks can predict long-distance subject-verb number agreement, even in the presence of other potential agreement triggers (often called attractors) intervening between the subject and verb. Gulordava et al. (2018) show that LSTMs trained on four different languages can accurately predict subject-verb agreement even when they are not trained specifically on that task. Futrell et al. (2019) show that surprisal scores of LSTMs (a measure of successful prediction) mirrored preferences of human participants on grammatical judgments task differentiating word-order alternations.

LSTMs therefore offer a powerful but convenient general-purpose learning mechanism for modelling human learning. Here we use LSTMs to process relatively short sequences: we

train models on artificially constructed inflectional paradigms which differ only in their i-complexity. The networks are trained and tested on wordforms plus suffixes. We then test them on (i) whether they are able to exploit the predictive information present in the lower i-complexity paradigm in order to generalise to novel forms, and (ii) whether the inflectional paradigm with lower i-complexity is learned faster.

2.3.1 Method

Target paradigms

We constructed two paradigms, which we used to test the effect of i-complexity in neural network and human learners. The basic paradigms both consisted of nine CVC nouns (*gob, tug, sov, kut, pid, tal, dar, ler, mip*), randomly paired with meanings for human participants (see Section 2.3.5 below). An additional nine nouns were used to test network generalization (*bor, ges, kiv, mas, nek, nap, lan, wib, log*) in Section 2.3.3 below. The small lexicon size allows the system to be learned with reasonable accuracy by human participants in a short experiment. The nouns were randomly allocated to three classes (for each run of the network, or each human participant), and each class was inflected for three numbers: singular, dual and plural. As in Seyfarth et al. (2014) noun class membership was not indicated by the semantics or phonology of the noun stem, but was rather defined only by different patterns of inflection. Inflectional markers were seven VC monosyllabic suffixes (*-op, -oc, -um, -ib, -el, -od, -at*). These inflectional markers were randomly allocated to cells in each paradigm (for each run of the network, or each human participant) such that both paradigms shared the same e-complexity value (1.14 bits) but differed in i-complexity. In the low i-complexity paradigm, the singular form of a word predicts the dual form, while in the high i-complexity paradigm it does not. Table 2.3 shows two example paradigms. In the low i-complexity

paradigm 2.3a, if a stem takes the marker -at in singular, then it takes -*oc* in dual; if a stem takes -*op* in singular, then it takes -*um* in the dual. In contrast, in the higher i-complexity paradigm 2.3b, there is not such regularity: nouns with -at in the singular take either -*oc* or -*um* in the dual. The i-complexity value for the low i-complexity language is 0.222 bits vs. 0.444 bits for the high i-complexity language. Note that the distinct plural forms in each paradigm serve to separate the three classes of nouns. Without distinct plural forms, the low i-complexity paradigm would have fewer classes and then high i-complexity paradigm.

	Singular	Dual	Plural
noun class 1	-at	-oc	-ib
noun class 2	-op	-um	-el
noun class 3	-at	-oc	-od

(a) low i-complexity paradigm

	Singular	Dual	Plural
noun class 1	-at	-oc	-ib
noun class 2	-op	-oc	-el
noun class 3	-at	-um	-od

(b) high i-complexity paradigm

Table 2.3: Example paradigm for low i-complexity (a) and high i-complexity (b) languages.

Learning in both computational and behavioral studies was staged – neural networks and human participants were first trained on the forms of the stems in singular and plural before being exposed to the forms in dual. This should increase the likelihood of learners exploiting the predictive relationships in the paradigm. The critical measure of learning was accuracy on dual forms, although we also report accuracy for singular and plural forms.

2.3.2 LSTM model

We implemented LSTM networks using the Keras package in Python (Chollet et al. 2015). Figure 2.1 presents a diagram of the network, which consists of an input layer, a hidden layer

of LSTM units, and an output layer fully connected to the hidden layer. In this task, the LSTM network takes as input a string (a sequence of characters) representing a noun stem and a number indicating the grammatical number of the object (1 for singular, 2 for dual and 3 for plural), e.g., the sequence *sov3* indicates the stem *sov* in plural. The input string is fed into the network incrementally. The network has seven output units, one for each of the inflectional suffixes in the target paradigms. The model is trained to generate the correct suffix for a stem, number sequence. Both input stem+number sequences and output suffixes were encoded as one-hot vectors. i.e., every character (and number) used in the language is represented as a vector of zeroes (with length equal to the total set of characters, 27, for the input, and the total set of suffixes, 7, for the output) with '1' in a different index uniquely identifying it. The algorithm used Stochastic Gradient Descent (SGD) to update the weights of the network during training.

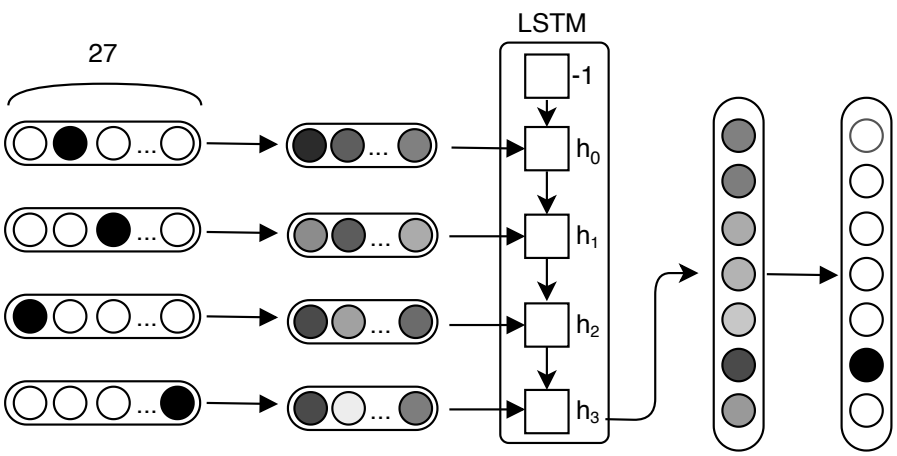


Figure 2.1: A diagram of the recurrent neural network: the input layer receives a string of four characters (stem + grammatical number), each coded as a one-hot vector of the length of the different characters used in the language (27). The input vectors are embedded and the embeddings are transferred to a hidden layer with 5-45 LSTM units. Output from the LSTM units (h_3) is then transferred to an output layer with seven options, representing the seven suffixes in the language. Using a softmax function, the output is converted to a one-hot vector, representing the suffix the network selected for this input.

2.3.3 Simulation Experiment 1 - generalizing to novel forms

We first tested whether low i-complexity facilitates solving the Cell Filling Problem (Ackerman, James P. Blevins, et al. 2009) in RNNs. We trained RNNs on nine stems in the full paradigm, and tested their accuracy at producing the correct dual suffix for nine additional stems, for which the network was trained on the singular and plural form but never the dual (i.e. for these 9 stems the network was required to generalise to the dual based on its representation of the full paradigm and its exposure to the singular and plural forms for those 9 stems). Stems were assigned randomly to one of the three noun classes in the language, such that each noun class included six stems (three fully trained, and three with dual held out). For each paradigm, the model was trained and tested on input-output pairs in three blocks, summarized in Table 2.4 below. In block 1, the network was trained and tested on singulars for all stems; in block 2 the networks was trained and tested on singulars and plurals for all stems; in block 3 the network was trained on singulars and plurals for all stems, plus duals for nine of the 18 stems. Finally, it was tested on the entire paradigm for all stems (i.e. including duals for the 9 stems where the dual form was held-out in training). Each block consisted of 300 epochs, each consisting of a single pass through the specified training set (randomized) with weights updated by backpropagation, followed by a pass through the specified test set (randomized). During testing, the network was given an input and had to generate an inflection. As noted above, weights in the network are tuned only during training, while during testing the network is tested on the same input without updating weights. Our results show performance in the testing phase.

Since we did not have a hypothesis regarding the appropriate network size for this task, we varied the network size, from 5-cells networks (582 parameters) to 45 (12,382 parameters), in increments of 5. For each network size, we conducted 100 runs of the model for each paradigm. In each run of the model, the initial weights were randomly generated, according

Block	Epochs	Training	Testing
1	300	all stems, singular only (18 items)	all stems, singular only (18 items)
2	300	all stems, singular and plural (36 items)	all stems, singular and plural (36 items)
3	300	9 fully-trained stems; 9 dual-held-out stems (45 items)	all stems, all numbers (54 items)

Table 2.4: Summary of training and testing blocks for RNNs in Simulation Experiment 1.

to a 'glorot_uniform' function (sampling from a uniform distribution in the range of $[-x, +x]$, where x is a function of the size of the network).

Results

Figure 2.2 presents the mean accuracy (averaged over all runs) on dual forms for dual-held-out stems by networks trained on the two paradigms. Networks trained on the low i-complexity paradigm achieved higher accuracy than networks trained on the high i-complexity paradigm across all sizes of the network; in other words, networks trained on the low i-complexity paradigm were more accurate in generalizing the paradigm to forms they were not trained on. Note that this is the case even though, in both the low i-complexity and high i-complexity paradigms, it is possible to infer the dual form of these nouns (untrained) from their plural forms (trained). For example, looking at the high i-complexity paradigm in Table 2.3, if the plural form of the stem *sov* is *sovib*, then it must be in noun class 1, and therefore its form in dual should be *sovum*. The predictive function of the plural thus allows networks trained on both paradigms to reach high accuracy, but there is nonetheless a clear advantage to the network trained on the low i-complexity language, where the singular provides an additional cue. In the Appendix we show results for networks trained only on the singular forms of the novel stems (rather than exposing them to both the singular and plural forms): under that training regime, the networks trained on the high i-complexity

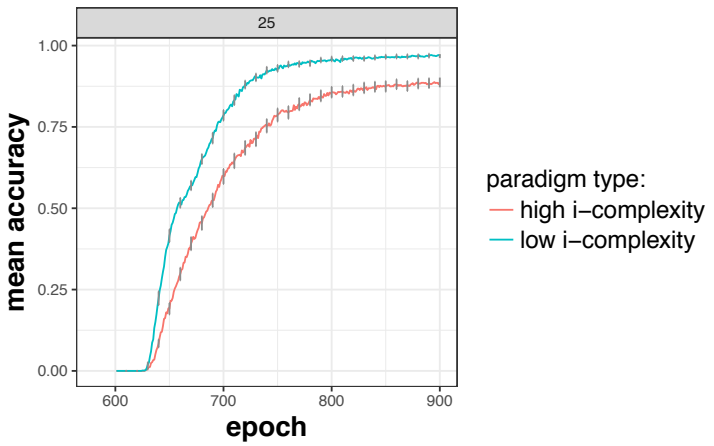
paradigm show accuracy of around 66% on generating duals for novel stems (chance level when guessing the more frequent dual suffix), while networks trained on the low i-complexity paradigm still show accuracy of almost 100% when generalizing to the dual (see Appendix Figure A.1).

These results show that for LSTM networks, low i-complexity facilitates solving the Paradigm Cell Filling Problem. This is in line with results from Seyfarth et al. (2014). As discussed above, learning a morphological system involves not only generalizing to completely novel stems, but also learning and retrieving forms for stems which are not yet robustly learned (e.g., due to low exposure frequency). We test this more general facilitative effect of low i-complexity in the next set of simulations.

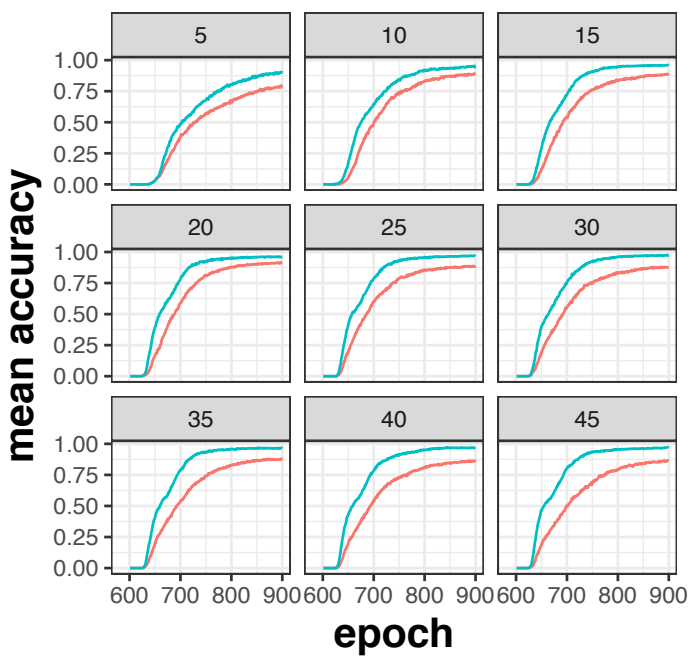
2.3.4 Simulation Experiment 2 - learning speed

The model

We tested LSTM networks with the same architecture, input and output representations, and parameters as described for Simulation Experiment 1. In this simulation, the language includes nine stems. The model was trained and tested in three blocks, summarized in Table 2.5. In the first block, the network was trained and tested on singulars for all stems; in block 2 the network was trained and tested on singulars and plurals for all stems; in block 3 the network was trained and tested on the entire paradigm for all stems, including the dual. Note that, unlike for Simulation Experiment 1, here the test set is always identical to the training set, i.e. we are not testing on the capacity of the network to generalize, but simply to learn the mapping from stem-number input to the appropriate affix. As before, we tested networks of different sizes, from 5-cells networks (542 parameters) to 45 (12,022 parameters), in increments of 5; we conducted 100 runs for each paradigm for each network



(a)



(b)

Figure 2.2: Average accuracy across all runs of the LSTM networks in generalizing to novel dual forms for the low i-complexity paradigm (blue) and the high i-complexity paradigm (red). (a) results for one network size (25 cells), with error bars indicating standard error every 10 epochs. (b) results for all the network sizes tested (facet titles give network size in number of cells). Note that the plots start at epoch 600, when the dual forms are introduced to the network (at the beginning of Block 3). In all cases accuracy is higher for the low i-complexity paradigm.

size, with initial weights of the network randomly generated for each run.

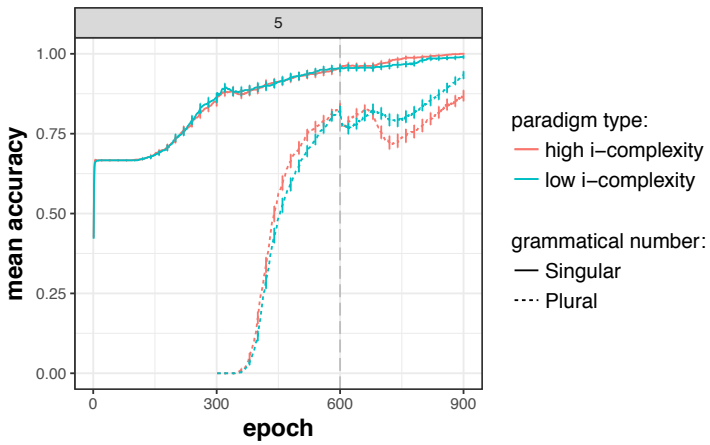
Block	Epochs	Training	Testing
1	300	all stems, singular only (9 items)	all stems, singular only (9 items)
2	300	all stems, singular and plural (18 items)	all stems, singular and plural (18 items)
3	300	all stems, all numbers (27 items)	all stems, all numbers (27 items)

Table 2.5: Summary of training and testing blocks for RNNs in Simulation Experiment 2.

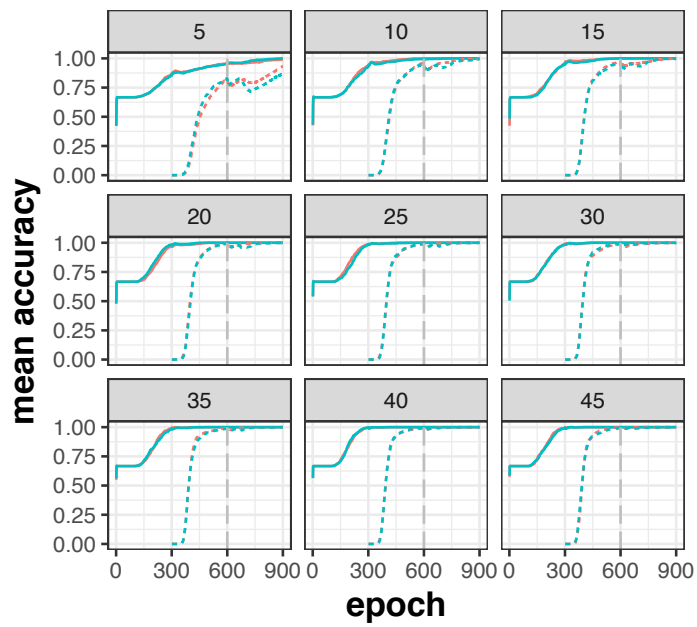
Results

Figure 2.3 presents the learning trajectories of the neural networks for the singular and plural forms in the low and high i-complexity paradigms. Networks of all sizes show similar learning trajectories and final accuracy levels across both paradigms, as expected since there is no difference between the paradigms for these forms. In all sizes of networks, except for size 5 (Figure 2.4a), the networks reach perfect learning of the singular and plural forms (accuracy of 1), and the singular and plural forms are fully learned (or near perfect learning) by the beginning of block 3 (epoch 600), when the dual forms are introduced to the network. These results verify that the networks are able to learn the singular and plural forms. This is necessary in order to exploit the predictive structure of the paradigms to better learn the dual. Interestingly, the smaller networks show a small decrement in accuracy of plural learning after epoch 600, the point at which the dual is introduced; networks of size 5 show some differences between high and low i-complexity in learning the plural forms after that point, with a more rapid recovery for networks trained on the low i-complexity paradigm.

Figure 2.4 presents the learning trajectories of the neural networks for the dual forms. Across all network sizes, the dual forms are learned faster in the low i-complexity paradigm. This difference in learning speed for low vs high i-complexity can also be seen in Figure 2.5,



(a)

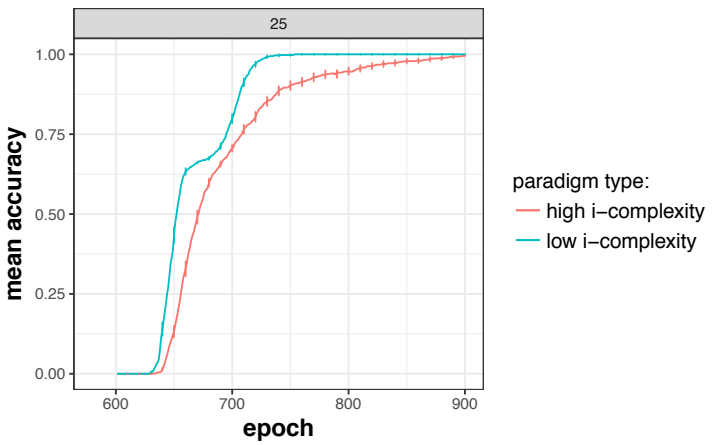


(b)

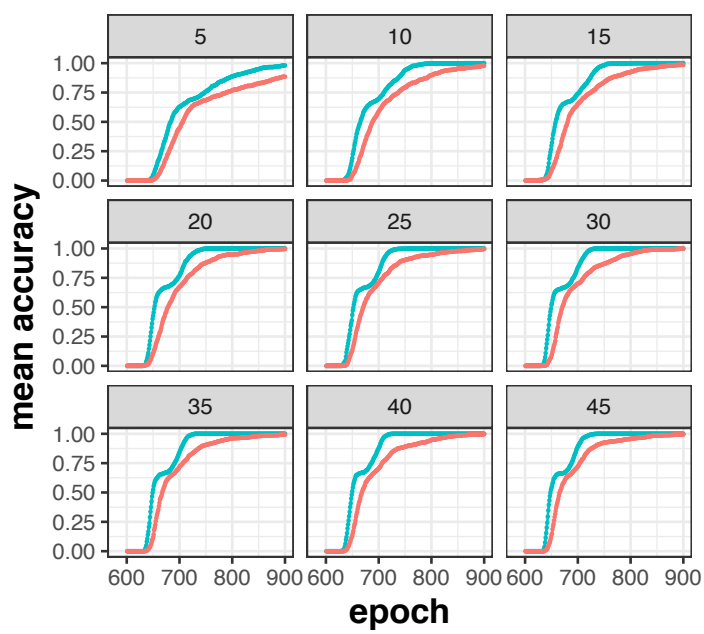
Figure 2.3: Network learning trajectories for singular and plural forms for high and low i-complexity paradigms. (a) results for one network size (5 cells), with error bars indicating standard error every 10 epochs, (b) results for all the network sizes tested (facet titles give network size in number of cells). Note that dashed lines for plural suffixes start at epoch 300 (block 2). Networks exposed to the high i-complexity language and low i-complexity language show similar performance.

showing for every network size the epoch in which the network reached perfect learning of the full paradigm. These results show that the predictive relationships in low i-complexity

paradigms facilitate learning in these networks beyond generalisation to entirely novel stems.



(a)



(b)

Figure 2.4: Network learning trajectories for dual forms for high and low i-complexity paradigms. (a) results for one network size (25 cells), with error bars indicating standard error every 10 epochs, (b) results for all the network sizes tested (facet titles give network size in number of cells). Note that the plots start at epoch 600, when the dual forms are introduced to the network (block 3).

Together, results from Simulation Experiments 1 and 2 show that for LSTM neural net-

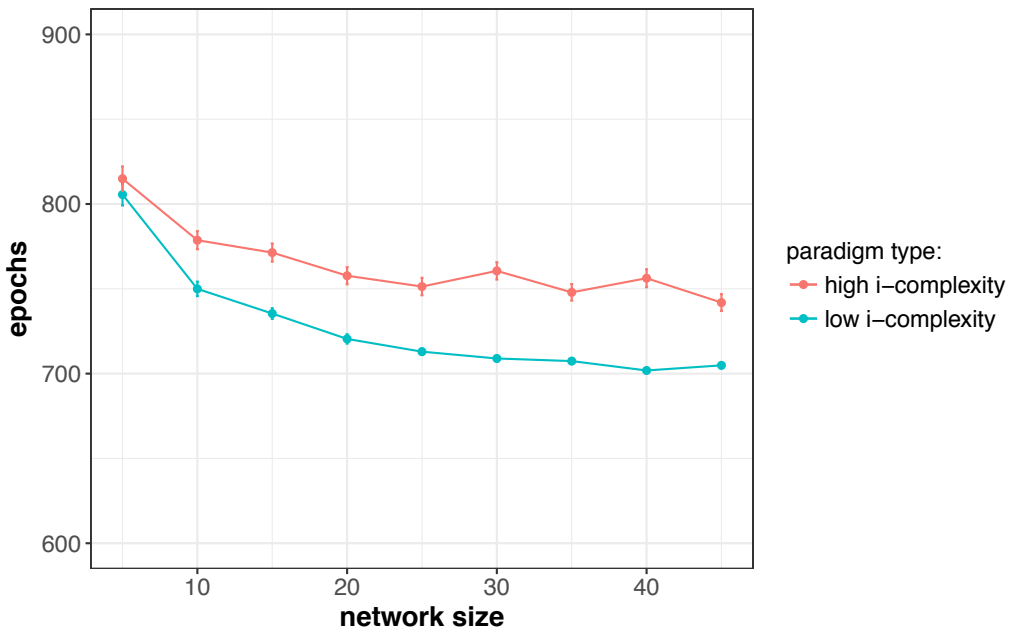


Figure 2.5: Number of training epochs required to reach perfect learning of the paradigm for each size of network.

works, the i-complexity measure is predictive of learning and generalisation when controlling for other factors, such as number of different markers, e-complexity, and inflection frequency. When trained on inflectional paradigms with low i-complexity, LSTM networks showed higher accuracy in generalizing to completely unseen dual forms, and more rapid learning of dual forms which were trained at low frequency. In principle then, and for at least one learning model, i-complexity influences generalization and learning of inflectional paradigms. This is consistent with the hypothesis that i-complexity –a measure of the predictive structure of a paradigm- reflects the learnability of inflectional paradigms. However, we are specifically interested in human learning, since it is biases in human learning that will shape human languages. While RNNs may mimic some features of human learning they cannot fully simulate human learning (e.g., Gulordava et al. 2018; Linzen et al. 2016). Below, we present a series of artificial language learning experiments to test the effects of i-complexity with human participants.

2.3.5 Experiment 1

As discussed above, Seyfarth et al. (2014) provide some evidence that human learners use the paradigmatic information captured by i-complexity to predict new forms. They found that, in a 2x3 paradigm, learners used the similarity between novel forms and trained forms to guess inflectional endings. In a slightly larger 3x3 paradigm, learners used this predictive information to guess low frequency inflectional endings. In the following experiments, we test whether i-complexity affects the speed with which a paradigm is learned, rather than the ability to generalize to completely novel forms. Participants were asked to learn labels for objects through trial and error, where these labels were drawn from one of the two artificially constructed paradigms described above. As in the network simulations, the i-complexity of the language was manipulated between subjects; the only difference between paradigms was the predictability of the dual form, which in the low-complexity language was predictable from the singular. Predictive relationship among forms in a paradigm can in principle facilitate learning of new (initially low-frequency) forms, and indeed neural networks showed precisely this benefit in our Simulation Experiment 2. If i-complexity also mediates human learning in this context, then participants in the low i-complexity condition should show faster learning and higher accuracy in inflecting nouns in the dual after learning the singular and plural forms.

Materials

Participants were trained on one of the two paradigms in Table 2.3 above. While our networks simulations did not involve any referents, here the nine stems referred to a set of simple objects (lemon, cow, tomato, bicycle, horse, clock, pigeon, mug and pear) depicted by photographs. Singular nouns corresponded to a single object, dual corresponded to two

of the objects, and plural ranged from 3-12 randomly (see Figure 2.6). Stems and suffixes were randomly paired with meanings for each participant. As before, stems were randomly allocated to classes. In addition, noun class membership was not conditioned on meanings; every noun class had one animate object (cow/pigeon/horse), one edible object (tomato/lemon/pear) and one other (clock/bicycle/mug).

Participants

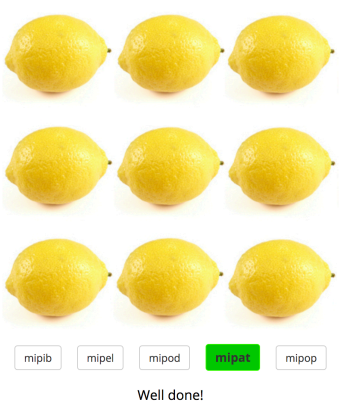
39 self-reported English speakers adult participants were recruited via Amazon's Mechanical Turk crowd-sourcing platform. They were paid $4.50 for their participation. Participants were allocated randomly to one of the two conditions (20 in the high i-complexity condition, 19 in the low i-complexity condition).

Procedure

On each trial, a picture was presented on the screen together with a set of possible stem + suffix labels, as in Figure 2.6a. Participants were asked to choose the correct label, and received feedback on their answer as in Figure 2.6b, 2.6b. The task was divided into 3 blocks of trials. In block 1 (36 trials), participants were exposed to the singular forms of all stems. In block 2 (72 trials) plural trials of all stems were introduced along with singulars. Finally, in the critical block 3 (108 trials), participants were exposed to all stems in all cells of the paradigm, including the dual. Each word form was presented four times in each block of trials. The different forms were randomly interspersed within each block.

(a)

(b)



(c)

Figure 2.6: Example plural trials. (a) a picture is presented and participants are asked to choose the correct label from a set of options. (b), (c) participants receive feedback on their answer, including the correct label. (b) negative feedback following trial shown in (a), (c) positive feedback following plural trial with a different number of objects.

Results

Figure 2.7 shows the mean accuracy with which participants chose the appropriate word form for singular, plural, and dual, as the experiment progressed trial by trial. Recall that we are particularly interested in how well participants learned the dual forms in block 3, after being trained on the singular and plural forms. Participants exposed to the high i-complexity language had mean accuracy in dual trials of M_H=0.58 (sd = 0.19) whereas in the low i-complexity condition mean accuracy was M_L=0.50 (sd=0.23).

To test the effect of i-complexity on production of dual forms, we ran a mixed effects logistic regression model predicting dual accuracy rates by condition (high vs. low i-complexity), accuracy on block 2, trial number, and their interactions as fixed effects.[4] Trial number was scaled and centered such that estimates for the effect of condition reflect the difference between conditions mid-way through block 3. Condition was sum-coded (high i-complexity = –1, low i-complexity = 1). We included participants' accuracy in block 2 as a way of controlling for general differences in learning ability. This is crucial, since participants in the high i-complexity condition were actually more accurate in blocks 1 and 2, despite no difference in the training participants had received at this point in the experiment.[5] Accuracy at block 2 was centered and scaled such that estimates for the effect of condition reflect the difference between conditions for participants with average accuracy in block 2. The model revealed a significant interaction between accuracy in block 2 and trial number (b=1.387, z=2.856, p=0.004), indicating that participants who learned the singulars and plurals better in block 2 learned the duals more rapidly in block 3. Results from the model also show significant effect for trial number (b=2.05, z=4.44, p<0.001), showing that participants'

[4]All models reported here were run using the lme4 package in R (Bates et al. 2014). All models include by-participant intercepts and random slopes for trial number.

[5]This was confirmed by a mixed-effects logistic regression model predicting accuracy in block 2 by condition and trial number (high vs. low i-complexity: b=-0.268, z=-2.198, p=0.028).

accuracy across conditions improved over trials. Crucially however, there was no significant effect of condition (b=0.11, z=0.72, p=0.471) on dual accuracy rates, and no significant interaction between condition and trial number (b=0.363, z=0.791, p=0.429).

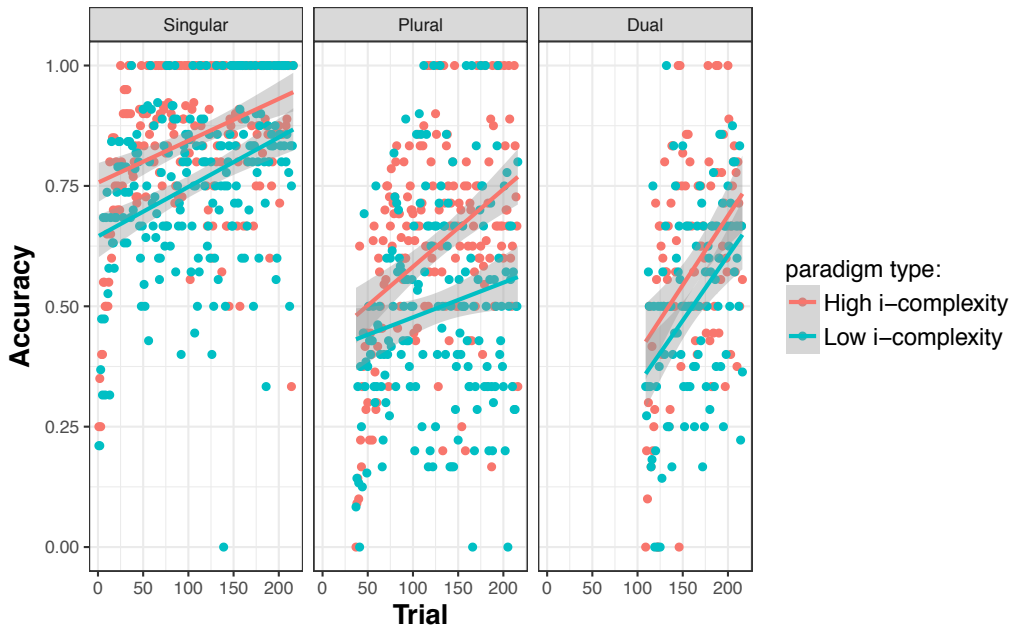


Figure 2.7: Mean accuracy by trial for singular, plural, and dual forms. Points indicates participants' mean accuracy scores in the low and high i-complexity conditions, with a regression line predicting accuracy by trial number for each grammatical number. Participants in the high i-complexity condition (unexpectedly) showed better accuracy in blocks 1 and 2. When controlling for accuracy in block 2, there was no significant effect of condition on dual learning.

Our results suggest that there is no learnability advantage for the low i-complexity paradigm; numerically our participants exposed to high i-complexity paradigms actually learned more quickly, but our statistical analysis shows that any apparent difference in dual accuracy across the two conditions in Figure 2.7 is in fact driven by individual differences in learning ability/attentiveness, rather than the structure of the target paradigm. This experiment therefore fails to confirm the predicted learning advantage for lower i-complexity paradigms.

There are at least three possible explanations for why we fail to find this predicted advantage

in our task. One possibility is that, for human learners, lower i-complexity provides some advantage in generalizing to completely novel words (Seyfarth et al. 2014), but does not continue to facilitate learning of forms already encountered, even in low frequency. Here, it is also possible to learn the forms through memorization, ignoring any predictive structure in the paradigm that might be helpful. Another possibility is that we don't see the effect i-complexity for methodological reasons; either because we don't have enough critical (dual) trials to reveal differences between the two conditions, or, because the predictive relationship from one form to another was not readily accessible to participants. In our task, the different forms of the nouns were randomly interspersed within each block, while learners might need a more explicit cue to the implicative structure of the paradigm. Studies in first language acquisition suggest that variation sets in the input to children, pairs of utterances that balance their overlap and change facilitate vocabulary growth (e.g., Brodsky and H. Waterfall 2007; Tal and Arnon 2018; H. R. Waterfall 2005). In a similar way, presenting the new forms in the language after presenting a familiar form of the same noun could make the predictive structure of the paradigm more apparent for learners. In Experiments 2-3, we therefore added more critical trials (a 4th block) and structured trials to highlight the predictive structure of the paradigms. In particular, we organized the trials in blocks 3-4 in pairs, so that dual trials always followed either plural or singular trials, with the same object. This design is parallel to the task Seyfarth et al. (2014) used to test generalization to novel forms, but note that here we are always presenting and testing on familiar nouns.

2.3.6 Experiment 2

Materials

All materials were identical to Experiment 1.

Participants

41 self-reported English speakers adult participants were recruited via Amazon's Mechanical Turk. They were paid \$6 for their participation. Participants were allocated randomly to one of the conditions (20 high i-complexity, 21 low i-complexity).

Procedure

The procedure was identical to Experiment 1 with two exceptions. First, the task included 4 blocks of trials, where block 4 was identical in its structure to block 3. Second, blocks 3 and 4 included all word forms in the paradigm, set up in pairs: singular or plural trial were always immediately followed by a trial with the same object in a different number, as illustrated in Figure 4.5. Critically, when the following trial was dual, the predictive information in the low i-complexity paradigm could be particularly helpful to learners. This arrangement resulted in three types of pairs: singular trial followed by a dual trial ('predictive trials'), plural followed by dual trial and singular followed by plural. The number of each type of pair was balanced and different pairs of trials were randomly interleaved.

Results

Figure 2.9 shows the mean accuracy with which participants chose the appropriate word form for singular, plural, and dual, as the experiment progressed trial by trial. Mean accuracy on the critical dual trials is higher for the low i-complexity condition (M_L=0.71, sd = 0.23) than for the high i-complexity condition(M_H=0.55, sd = 0.26). However, as in Experiment 1, differences between conditions already appeared in blocks 1 and 2, despite the fact that

(a) (b)

Figure 2.8: Example of two successive trials in blocks 3 and 4 in Experiment 2. (a) trial n, in which participant is asked to choose the correct form describing a mug in singular, (b) trial $n + 1$, in which participant is asked to choose the correct label for the same object in dual.

the input languages are identical up to this point.[6] We ran a mixed-effects logistic regression model predicting dual accuracy rates by condition (high vs. low i-complexity), accuracy on block 2, trial number and their interactions as fixed effects, with fixed effects coded as in Experiment 1. Results again revealed a significant effect of trial number (b=1.45, z=7.81, p<0.001) and a significant interaction between accuracy in block 2 and trial number (b=0.66, z=3.441, p<0.001), but no significant effect of condition (b=0.141, z=0.806, p=0.42), and no condition by trial number interaction (b=0.223, z=1.25, p=0.211). These results hold when looking only at predictive trials (dual trials following singular trials): the interaction of accuracy in block 2 with trial number is significant (b=0.652, z= 3.1, p=0.0019) and so is the effect of trial number (b=1.09, z=5.7, p<0.001), but there is no main effect of condition (b=0.142, z=0.87, p=0.383), and no interaction between condition and trial number (b=0.191, z=1.03, p=0.301). To summarize, lower i-complexity did not lead to a learning

[6]This was confirmed by a mixed-effects logistic regression model predicting accuracy in block 2 by condition and trial number (high vs. low i-complexity: b=0.314, z=2.395, p=0.016). In this case, participants allocated to the low i-complexity condition were significantly better in block 2.

advantage when controlling for participants' earlier learning, not even in predictive trials, where the singular form which predicts the dual is easily accessible.

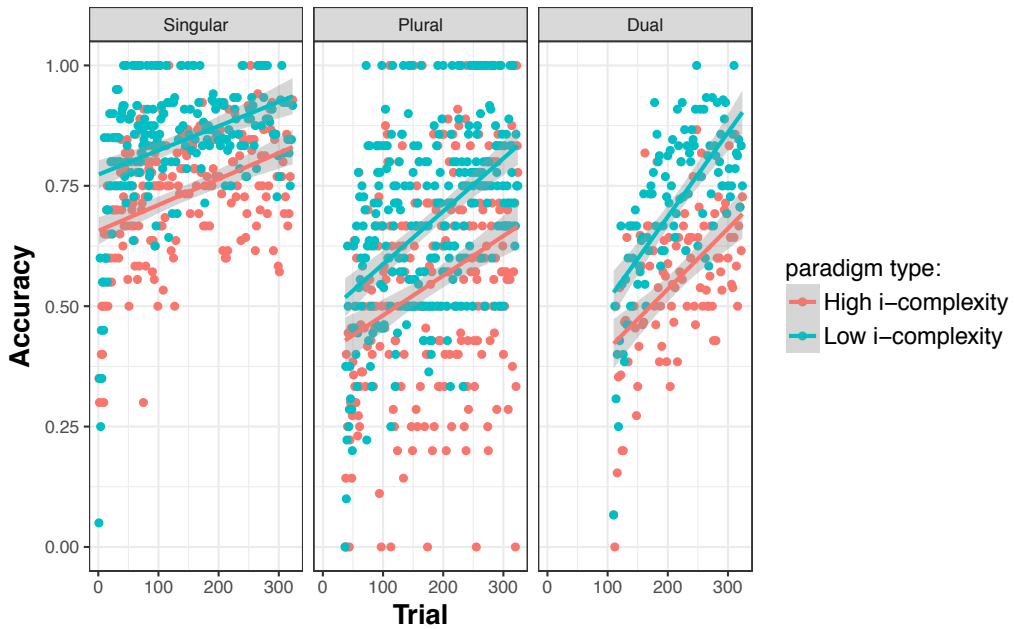


Figure 2.9: Mean accuracy by trial for singular, plural, and dual forms in Experiment 2 (with predictive trials). Points indicates participants' mean accuracy scores in the low and high i-complexity conditions, with a regression line predicting accuracy by trial number for each grammatical number. Participants in the low i-complexity condition (unexpectedly) showed better accuracy in blocks 1 and 2. When controlling for accuracy in block 2, there was no significant effect of condition on dual learning.

2.3.7 Experiment 3

While Experiments 1 and 2 are consistent in showing no advantage for low i-complexity, our confidence in this conclusion was reduced by the substantial inter-individual differences in our sample; in both Experiments we found substantial differences between conditions in learning accuracy in block 2, before the conditions had diverged. In Experiment 3 we replicate Experiment 2 with more participants in order to reduce the likelihood of obtaining samples with accidental differences between the two conditions; since Experiment 3 uses an

identical method to Experiment 2, a combined analysis (total N=100) also becomes possible.

Participants

59 self-reported English speakers adult participants were recruited via Amazon's Mechanical Turk. They were paid $6 for their participation. Participants were allocated randomly to one of the conditions (30 high i-complexity, 29 low i-complexity).

Materials and Procedure

This experiment was identical to Experiment 2.

Results

Figure 2.10 shows the mean accuracy with which participants chose the appropriate word form for singular, plural, and dual, as the experiment progressed trial by trial. In this case, learners in both conditions were more balanced with respect to their general ability to learn in the task. Mean accuracy in dual trials was higher in the low i-complexity condition (M_L=0.69, sd=0.25) than in the high i-complexity condition (M_H=0.55, sd=0.25). We used a mixed-effects logistic regression model, as in Experiments 1-2, to predict dual accuracy based on condition (high vs. low i-complexity), accuracy on block 2, and trial number. The model revealed a significant effect of condition (b=0.328, z=2.208, p=0.027), as well as a significant interaction between accuracy in block 2 and trial number (b=0.85, z=5.29, p<0.001). The interaction between condition and trial number was not significant (b=0.189, z=1.26, p=0.205). This pattern of results suggests that, as in Experiments 1 and 2, participants who showed better learning in block 2 also showed faster learning of the dual forms. However, in addition, participants in the low i-complexity group were better at learn-

ing the dual forms; in other words, there was a learning advantage for participants in the low i-complexity condition, when controlling for individual differences in learning abilities. These effects also hold when looking only at predictive trials: the main effect of condition is significant (b=0.33, z=2.00, p=0.045), although the interaction of condition with trial number is not (b=0.12, z=0.722, p=0.47).

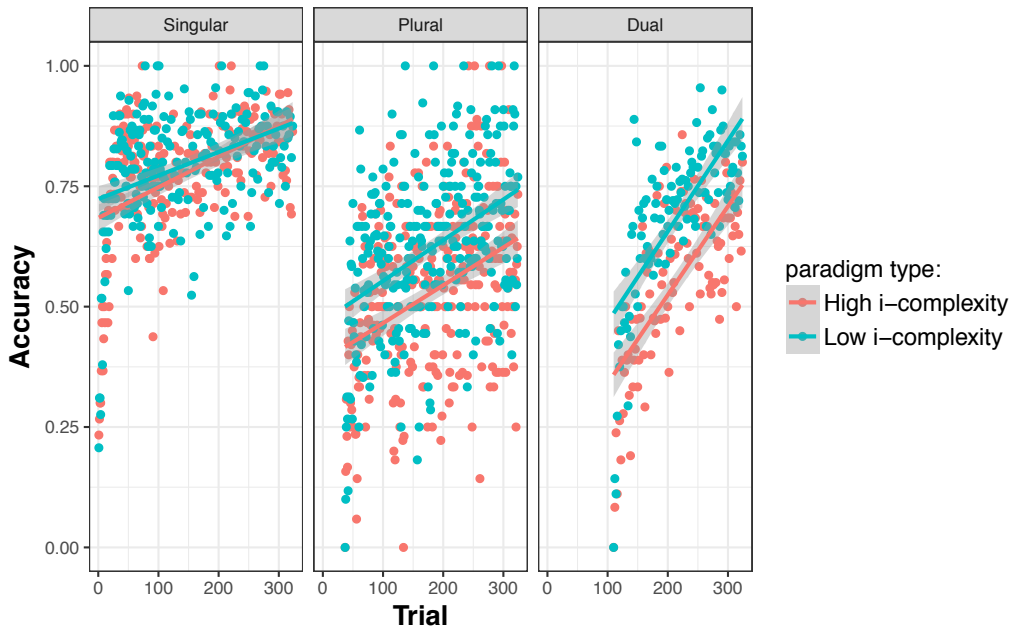


Figure 2.10: Mean accuracy by trial for singular, plural, and dual forms in Experiment 3 (replication with predictive trials). Points indicates participants' mean accuracy scores in the low and high i-complexity conditions, with a regression line predicting accuracy by trial number for each grammatical number. Participants were well matched in blocks 1 and 2. When controlling for accuracy in block 2, participants in the low i-complexity condition were better in learning the dual forms.

Since Experiments 2 and 3 were identical, we also ran a combined analysis including both data sets. We ran a mixed-effects logistic regression model, as before, predicting accuracy on dual trials from condition, experiment, trial number, and accuracy in block 2. The model revealed a significant effect of condition (b=0.236, z=2.012, p=0.044) as well as a significant interaction between accuracy in block 2 and trial number (b=0.7625, z=5.924, p<0.001). The

interaction of condition and trial number was not significant (b=0.214, z=1.81, p=0.070), and there was no significant effect of experiment (b=0.04, z=0.342, p=0.732). The results suggest that, after correcting for the random imbalance in learners across the two conditions, we see some evidence of an advantage for the low i-complexity language, in higher overall performance.

To test whether the effect of condition found in Experiments 2 and 3 was directly related to the presence of predictive trials, we also ran an additional analysis comparing data across all three experiments. Recall that in the predictive trials design used in Experiments 2-3, participants were asked for the dual form immediately after getting the singular form in the previous trial. If learners only make use of the predictive information when they have ready access to a predictive form, these trials should facilitate learning of the dual forms, but only in the low i-complexity conditions of Experiments 2 and 3, not Experiment 1. We ran a mixed-effects logistic regression predicting accuracy on dual trials from experiment, condition, accuracy in block 2, and trial number. Coding for condition and trial number was the same as in previous analyses. Experiment was sum coded (Experiment 1 = –1, Experiments 2 and 3 = 1). Trials from block 4 in Experiments 2 and 3 were excluded from the analysis, to match number of blocks and trials across the three Experiments. A benefit for predictive trials should manifest as an interaction between experiment and condition in this model. The model revealed a significant interaction between trial number and experiment (b=0.265, z= 2.63, p=0.008), but no significant interaction between experiment and condition (b=0.067, z=0.716, p=0.474) and no significant effect of condition (b=0.143, z=1.527, p=0.12). This suggests that the predictive trials design was actually beneficial for participants in both conditions; being exposed to one form of a noun helped participants to subsequently retrieve the dual, regardless of the i-complexity of the paradigm. This is in line with studies showing that repetition of stems in different contexts (sometimes called overlap) facilitates word

learning in first language acquisition (Brodsky and H. Waterfall 2007; Tal and Arnon 2018; H. R. Waterfall 2005). However, there is no evidence that the predictive trials provided any particular advantage for participants in the low i-complexity condition, for example by making the predictive link between forms in the low i-complexity paradigm more apparent to learners.

At this point, our results show only quite weak evidence for effects of i-complexity on learning inflectional paradigms. On the one hand, we see a significant effect of i-complexity on learning in one of our experiments (Experiment 3) and in the combined analysis of 100 participants across Experiments 2 and 3, in which we used the same method, namely the predictive trials in blocks 3 and 4 of the experiments. However, Experiments 1 and 2 when considered independently do not show a significant effect of i-complexity on learning, and a combined analysis of all three experiments (combined N=139) does not show a significant effect of i-complexity or that the predictive trials method interacts with i-complexity.

2.4 Testing the impact of e-complexity on paradigm learning

Given the rather weak evidence of an effect of i-complexity for human learners, we attempted to test whether another measure of paradigm complexity (e-complexity) had a more robust effect; recall that natural language paradigms differ quite substantially on e-complexity. Using the simulation and experimental methods above, we manipulate e-complexity to test its effect on paradigm learning in RNNs and human participants and to compare the effect this measure has on learning with that of i-complexity. The paradigms used in Simulation Experiments 1-2 and Experiments 1-3 varied in i-complexity but had low e-complexity.

We constructed a third inflectional paradigm similar in size and design to the ones used above, with high e-complexity and low i-complexity (Table 2.6) and tested how well RNNs and human participants learned the forms in the paradigm. In Simulation Experiment 3 we compare results from RNNs learning the paradigm with results from Simulation Experiment 2. In Experiment 4 we compare results from human participants learning the high e-complexity paradigm with combined data from Experiments 2 and 3. This allows us to compare the effect that the e-complexity of the paradigm has on its learnability with the effect of i-complexity.

2.4.1 High e-complexity paradigm

In this Experiment, RNNs and participants are trained and tested on an inflectional paradigm of an artificial language. Higher e-complexity is reflected by having more variants to mark the dual forms. We use the same paradigm in the simulation experiment with RNNs and in the experiment with human participants.

	singular	**Dual**	**Plural**
noun class 1	-at	-oc	-at
noun class 2	-op	-um	-el
noun class 3	-at	-ib	-od

Table 2.6: Example paradigm for high e-complexity language. This paradigm is of the same size as in previous experiments (3 noun classes, 3 grammatical numbers, 7 inflections), but with higher *e-complexity*.

The high e-complexity paradigm (Table 2.6) has higher e-complexity than the paradigms shown in Table 2.3, 1.36 vs. 1.14 bits, while its overall i-complexity is equal to that of the low i-complexity paradigm in Table 2.3a, 0.222 bits. The higher e-complexity is due to an additional variant to mark the dual; we keep the overall number of inflections constant by re-using one marker from the singular in the plural (in this example, -at).

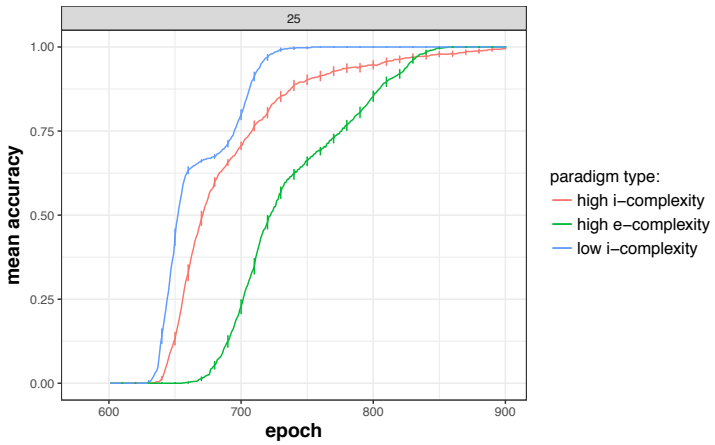
2.4.2 Simulation Experiment 3

The model

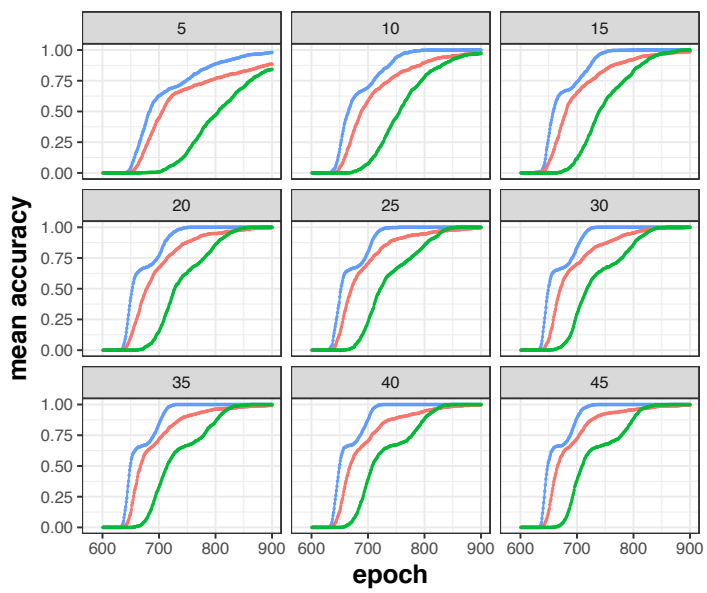
We tested LSTM networks with the same architecture, input encoding and parameters as described for Simulation Experiments 1 and 2. Training and testing the network on the forms according to the paradigm in Table 2.6 followed the same procedure as in Simulation Experiment 2, summarized in Table 2.5. As before, we tested networks of different sizes, from 5-cells networks (542 parameters) to 45 (12,022 parameters), in increments of 5; we conducted 100 runs for each paradigm for each network size, with initial weights of the network randomly generated for each run.

Results

Figure 2.11 presents the learning trajectories of the neural networks for the dual forms, comparing the results from networks trained on the paradigm with new high e-complexity with results from Simulation Experiment 2. We designate the two paradigms from Table 4 as low i-complexity and high i-complexity; note that both these paradigms have lower e-complexity compared to the new paradigm in Table 2.6; our high e-complexity paradigm has the same i-complexity as the low i-complexity paradigm in Table 2.3a. Across all network sizes, the dual forms in the high e-complexity paradigm are learned slower than both the low i-complexity and high i-complexity paradigms. This difference in learning speed can also be seen in Figure 2.12, showing for every network size the epoch in which the network reached perfect learning of the full paradigm.



(a)



(b)

Figure 2.11: Average accuracy across all runs of the LSTM networks in generalizing to novel dual forms for the new high e-complexity paradigm (green) compared with results from the low i-complexity paradigm (blue) and the high i-complexity paradigm (red) from Simulation Experiment 2. (a) results for one network size (25 cells), with error bars indicating standard error every 10 epochs. (b) results for all the network sizes tested (facet titles give network size in number of cells). Note that the plots start at epoch 600, when the dual forms are introduced to the network (at the beginning of Block 3). In all cases accuracy for the high e-complexity paradigm is lower than for both other paradigms.

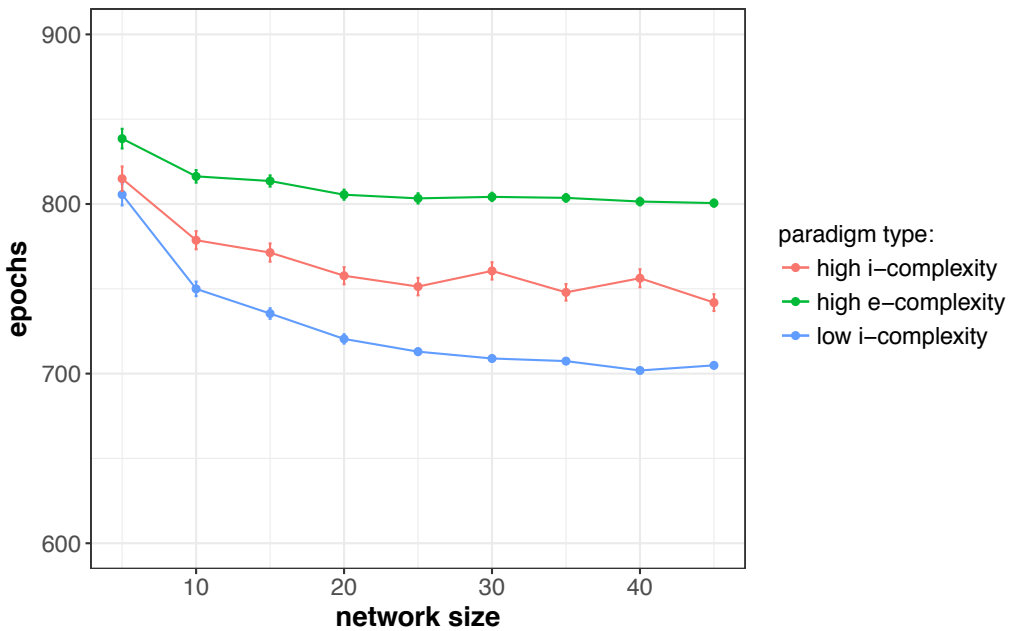


Figure 2.12: Number of training epochs required to reach perfect learning of the paradigm for each size of network.

2.4.3 Experiment 4

The results from Simulation Experiment 4 suggest that for LSTMs e-complexity has more of an effect on learning than i-complexity, Next, we test the effect of e-complexity on learning inflectional paradigms in human learners, and compare it with the effect of i-complexity. We use the combined data from Experiments 2-3 to represent paradigms with low e-complexity but varying i-complexity.

Participants

50 self-reported English speakers adult participants were recruited via Amazon's Mechanical Turk. They were paid $6 for their participation. All participants were allocated to the high e-complexity condition.

Materials and Procedure

This experiment was identical to Experiments 2 and 3, with the sole difference of training participants on an inflectional paradigm with high e-complexity (Table 2.6).

Results

Figure 2.13 shows the mean accuracy with which participants chose the appropriate word form for singular, plural, and dual, as the experiment progressed trial by trial. Results from this experiment are compared with data from Experiments 2 and 3, where participants were exposed to paradigms with low e-complexity and either low or high i-complexity. Participants exposed to the high e-complexity language had mean accuracy in dual trials of M_{highe}=0.45 (sd = 0.25) whereas in the low e-complexity paradigms participants mean accuracy was M_{low}=0.69 (sd=0.24) (in the low i-complexity condition) and M_{highi}=0.56 (sd=0.26) (in the high i-complexity condition).

There was no significant differences between conditions in blocks 1 and 2, in which the linguistic input did not differ in terms of i- and e-complexity across conditions up until block 3, where the dual forms were introduced.[7] To test the effect of e-complexity on production of dual forms and to compare it with the effect of i-complexity, we ran a mixed-effects logistic regression model predicting dual accuracy rates by condition (low i-complexity [with low e-complexity], high i-complexity [with low e-complexity], and high e-complexity [with low i-complexity]), accuracy on block 2, trial number and their interactions as fixed effects. Condition was dummy-coded with high e-complexity as the reference level, to compare the other two conditions to it. Other fixed effects were coded as in Experiments 1 and 2. The

[7]This was confirmed by a mixed-effects logistic regression models predicting accuracy in block 2 by condition and trial number (high vs. low e-complexity: b=0.167, z=0.97, p=0.33; high i-complexity vs. high e-complexity: b=-0.29, z=-1.7, p=0.088).

results of this experiment closely mirror the findings of Simulation Experiment 3. Results revealed a significant effect of e-complexity, with higher accuracy for the low i-complexity condition than the high e-complexity condition (b=1.502, z=7.3, p<0.001), and significant interaction between e-complexity and trial number (b=0.67, z=3.21, p=0.0013), suggesting that participants in the low i-complexity, low e-complexity condition improved faster in learning the dual forms than participants learning the new high e-complexity, low i-complexity paradigm. Accuracy in the high e-complexity condition was also significantly lower than in the high i-complexity, low e-complexity condition (b=0.99, z=4.9, p<0.001), but in this case there was no significant interaction between condition and trial number (b=0.299, z=1.46, p=0.14). As seen in earlier experiments there were also significant effects of trial number (b=1.04, z=7.44, p<0.001), accuracy in block 2 (b=1.3, z=8.26, p<0.001), and a significant interaction between accuracy in block 2 and trial number (b=0.52, z=3.21, p=0.0013).

To summarise, the results of experiments with both neural network and human participants suggest that when comparing the effect of e-complexity with the effect of i-complexity on learning, e-complexity has a stronger effect; a paradigm with high e-complexity but low i-complexity is learned more slowly than paradigms with lower e-complexity, regardless of whether those paradigms have the same or higher i-complexity.

2.5 Discussion

We used neural network simulations and behavioral experiments with humans to test whether the learnability of artificially-constructed inflectional paradigms was predicted by i-complexity. This information-theoretic measure of complexity, proposed by Ackerman and Malouf (2013), captures the extent to which forms in a paradigm predict each other. We tested whether a paradigm in which dual forms could be predicted from singular forms (lower i-complexity)

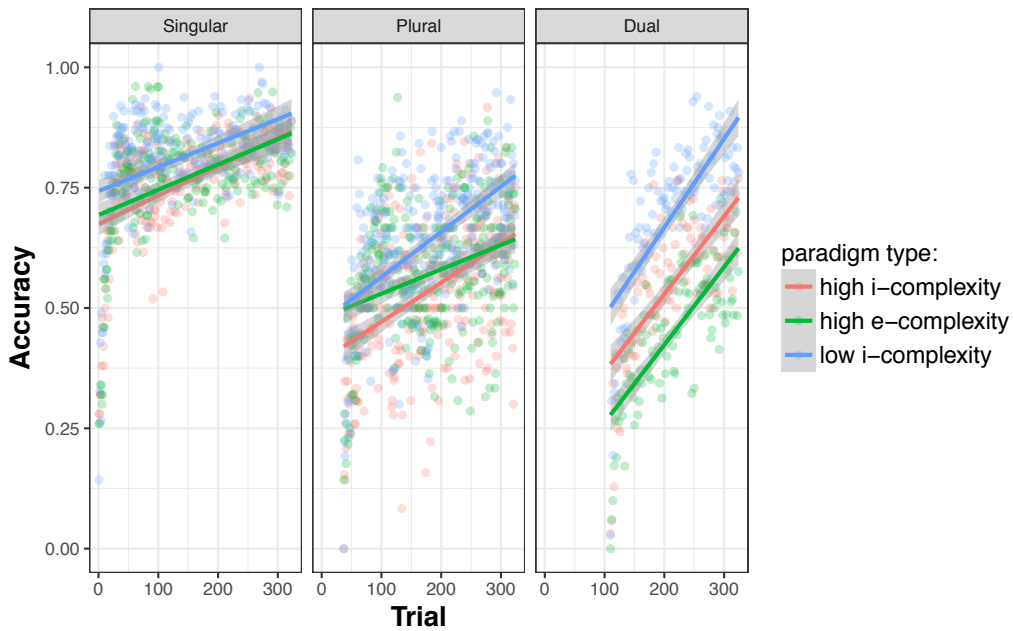


Figure 2.13: Mean accuracy by trial for singular, plural, and dual forms of participants exposed to the high e-complexity paradigm (in green) compared with results from Experiments 2 and 3 where participants were exposed to paradigms with low e-complexity and either low or high i-complexity. When controlling for accuracy in block 2, participants in the high e-complexity condition were worse in learning the dual forms than both low e-complexity paradigms.

was learned faster than a paradigm in which the dual is less well predicted from the singular (higher i-complexity), holding other measures of complexity (e.g., size of the paradigm and number of overall markers) constant. In contrast with previous work, we tested not just the ability to generalize to completely novel forms, but also the speed and overall accuracy with which the paradigms were learned. Results from network simulations with LSTMs showed that low i-complexity was advantageous both for generalizing the paradigm to novel forms (i.e., solving the Cell Filling Problem, Ackerman, James P. Blevins, et al. (2009)) and early learning of new forms that have not yet been robustly acquired (i.e., through memorization). These findings show that for an ideal learner, such as an LSTM neural network, the implicative structure of a morphological paradigm, captured by low values of i-complexity,

facilitates learning. However, since we are interested here in whether i-complexity shapes cross-linguistic variation in inflectional paradigms through its effect on learnability, it is crucial to verify the effect of i-complexity on human learners.

In previous work, Seyfarth et al. (2014) found evidence that learners use predictive structure to generalize to novel stems. In larger paradigms, with three inflectional classes each expressing three number features, predictive structure was at least partly over-ridden by the effect of inflection frequency: learners only used predictive structure to generate low frequency inflectional forms. Here, we tested whether lower i-complexity paradigms also led to better or faster learning. In Experiment 1, we found no evidence of a learning advantage for our lower i-complexity paradigm when controlling for general differences in learning ability. In Experiment 2, we attempted to reveal an effect of i-complexity by adding more critical trials, and by structuring trials such that a predictive form presented on one trial was followed the cell it predicted on the subsequent trial. In our case, this meant dual trials immediately followed singular trials for the same noun, a trial structure similar in spirit to that used by Seyfarth et al. (2014); structuring the task in this way in principle makes the predictive link between forms in the paradigm more apparent. As in Experiment 1 however, no difference between conditions was found after controlling for general learning ability. In Experiment 3 we replicated this with a larger sample size, and here found some evidence that participants were better-able to learn lower i-complexity paradigms. These results held in a combined analysis of data from Experiments 2 and 3 together. An analysis across all experiments revealed however that participants in both conditions benefited from our constrained ordering of trials. This suggests that there is no critical difference between the experiments with or without the predictive trials ordering in terms of the effect of i-complexity on learning. Further, in this combined analysis there was no evidence that i-complexity modulated learning. Across all three experiments, our results therefore suggest that implicative structure in

paradigms provides at best only a weak benefit for human learners.

That said, it is worth noting that the differences in i-complexity between our low- and high-complexity paradigms were not very large – the difference is 0.222 bits. It could be that larger differences in i-complexity values would reveal a larger effect on learning. However even this difference corresponds to complete predictability of the dual given the singular in the low complexity paradigm, vs at best 66% predictability in the high complexity paradigm. Testing more extreme values of i-complexity is in principle possible, but would necessitate training participants on much larger inflectional paradigms.

Alternatively, it may simply be that implicative structure is used for generalization to completely novel items, but not once learners have encountered items, even if with low frequency. However, this seems an unlikely explanation as that would presume perfect memory of which items have been encountered. Furthermore, as Seyfarth et al. (2014) show in their study, even when generalizing to novel items the effect of the predictive structure is secondary to inflection frequency (a feature better described with measures such as e-complexity). If so, the role of i-complexity in shaping natural language paradigms might be relatively weak. For example, measures of morphological complexity which take into account number of inflectional classes and frequency of inflectional forms, e.g., our measure for e-complexity, may in fact play a larger role.

In Simulation Experiment 3 and Experiment 4, we test the effect of e-complexity on learning inflectional paradigms to calibrate our understanding of the size of the i-complexity effect by seeing how it compares to an effect of e-complexity. We trained and tested RNNs and human participants on an inflectional paradigm with high e-complexity and low i-complexity. Results from both the simulation with RNNs and the experiment with human participants show a stronger effect on learning of e-complexity than i-complexity; our paradigm with high e-complexity was learned more slowly than the other paradigms we tested, even though

it had low i-complexity. These results show that e-complexity is a better predictor of the learnability of an inflectional paradigm. Therefore, the e-complexity or a combination of predictors of learning, rather than i-complexity alone, should be used to explain how learning biases shape morphological paradigms in natural languages.

Our results show that the measure proposed here for e-complexity is a more robust predictor of morphological learning than Ackerman and Malouf (2013)'s i-complexity. However, previous studies suggest that there are differences between adult and children learning (e.g., Culbertson and Elissa L. Newport 2015; Carla L. Hudson Kam and Elissa L. Newport 2005; Carla L Hudson Kam and Elissa L Newport 2009). Since we did not test children learning in our study, it may be that either there is a difference between L1 and L2 learning with respect to the effect of i-complexity, or that e-complexity is more dominant also in first learning acquisition. Future studies should look at the effect of i-complexity in children learning of morphological paradigms.

Furthermore, we show only weak evidence that differences in the reliability of analogy based on distributional cues affects the learning of morphological paradigms. It is however important to note that our results say nothing about using e.g. semantic or phonological similarities to generalize the paradigm to novel words. In the artificial language used here, class membership of lexical items was not conditioned on semantics or phonology; this was crucial to directly test the effect of the predictive structure on learning the inflectional forms independently from any reliance on other cues in predicting inflectional class membership (e.g., Culbertson, Gagliardi, et al. 2017; Frigo and McDonald 1998; L. A. Gerken et al. 2009; L. Gerken et al. 2005). Therefore, our results do not contradict evidence showing that learners use phonological similarities to generalize to new inflections (e.g., Ambridge 2010; Milin, Keuleers, et al. 2011); indeed, one possible future avenue of research is to test whether semantic or phonological cues interacting with distributional information might reveal advantages

for low i-complexity paradigms.

2.6 Conclusions

This paper was set to test the hypothesis that i-complexity predicts the learnability of morphological paradigms; an hypothesis arising from Ackerman and Malouf (2013)'s low conditional entropy conjecture. Using artificial language methods, we tested whether paradigms with low vs. high i-complexity are easier to learn, both with RNNs and with human participants. Results from this study show that while low i-complexity was shown to be beneficial for LSTM neural networks, it is not a strong predictor of learning in human learners; manipulating e-complexity shows a larger effect on learning both in LSTM networks and in human participants, a result which we replicate elsewhere (Johnson, Gao, et al. 2021).

Together, these results challenge the hypothesis that i-complexity has a role in shaping paradigms in natural languages. The mismatch between the effect of i-complexity on learning in LSTM networks and in human participants calls for future studies on the different mechanisms of these two learners in learning morphological paradigms. This is of high relevance since more and more studies are proposing ways of measuring morphological complexity and modelling morphological learning using neural networks (e.g. Cotterell, Kirov, Hulden, et al. 2019; Elsner et al. 2019; Malouf 2017; Marzi et al. 2018).

Part II

I-complexity and e-complexity and their effects on the learnability of morphological systems

Author Contributions

The following paper has been submitted to *Journal of Language Modelling*. The version included here is the submitted version. I conceived and designed the experiments and the simulations, conducted the analysis and wrote the paper; Kexin Gao collected the data and performed the simulations; Kenny Smith, Jennifer Culbertson and Hugh Rabagliati provided advice on the design of the experiments and data analysis, and commented on the paper.

Chapter 3

Investigating the effects of i-complexity and e-complexity on the learnability of morphological systems

3.1 Abstract

Research on cross-linguistic differences in morphological paradigms reveals a wide range of variation on many dimensions, including the number of categories expressed, the number of unique forms, and the number of inflectional classes. However, in an influential paper, Ackerman and Malouf (2013) argue that there is one dimension on which languages do not differ widely: in predictive structure. Predictive structure in a paradigm describes the extent to which forms predict each other, called i-complexity. Ackerman and Malouf (2013) show that although languages differ according to measure of surface paradigm complexity, called e-complexity, they tend to have low i-complexity. They conclude that morphological paradigms have evolved under a pressure for low i-complexity. Here, we evaluate the hypothesis that language learners are more sensitive to i-complexity than e-complexity by testing how well paradigms which differ on only these dimensions are learned. This could result in the typological findings Ackerman and Malouf (2013) report if even paradigms with very high e-complexity are relatively easy to learn so long as they have low i-complexity.

First, we summarize recent work by Johnson, Culbertson, et al. (2020) suggesting that both neural networks and human learners may actually be more sensitive to e-complexity than i-complexity. Then we build on this work, reporting a series of experiments which confirm that indeed, across a range of paradigms that vary in either e- or i-complexity, neural networks (LSTMs) are sensitive to both, but show a larger effect of e-complexity (and other measures associated with size and diversity of forms). In human learners, we fail to find any effect of i-complexity on learning at all. Finally, we analyse a large number of randomly generated paradigms and show that e- and i-complexity are negatively correlated: paradigms with high e- complexity necessarily show low i-complexity. We discuss what these findings might mean for Ackerman & Malouf's hypothesis, as well as the role of ease of learning versus generalization to novel forms in the evolution of paradigms.

Keywords: morphological complexity; learning; neural networks; typology

3.2 Introduction

Languages differ widely in their morphological systems, including substantial variation in their inflectional paradigms; some languages do not use morphology to mark grammatical information at all (e.g., Mandarin) whereas others make use of inflectional morphology to mark dozens of grammatical functions (e.g., Arabic). Intuitively, this kind of variation should have an effect on how easy or difficult it is to learn a morphological system—the more inflected forms for each lexeme there are, the more difficult learning should be. Indeed, using the size of an inflectional paradigm is a common method for measuring morphological complexity, for example by counting the number of potential inflections a verb or a noun can be marked with (e.g., Bickel and Nichols 2013; Shosted 2006). In addition to number of inflectional categories, the size of a morphological system is also impacted by the number of

inflection classes, i.e., different realizations for the same morphosyntactic or morphosemantic distinction across groups of lexemes (Aronoff 1994; Greville G. Corbett 2009), which has also been claimed to be a source of complexity in morphological systems (e.g., Ackerman and Malouf 2013; Baerman et al. 2010). These aspects of morphological complexity, which pertain to the size of a morphological system, are all referred to as enumerative complexity or e-complexity (e.g., Ackerman and Malouf 2013; Meinhardt et al. 2019).

Recently, another measure of the complexity of morphological paradigms has been suggested, referred to as integrative complexity, or i-complexity. I-complexity refers to the organization of the inflected forms in the paradigm and the relations between the forms that such organization generates; in paradigms with low i-complexity, forms are predictive of one another (e.g., Ackerman and Malouf 2013; James P. Blevins 2006). Proponents of this measure suggest that i-complexity reflects the difficulty speakers face in generating forms they have not previously encountered, based on known forms of the same lexeme (the Paradigm Cell Filling Problem, Ackerman and Malouf 2013; Ackerman and Malouf 2015.Predictive structure in a morphological system can be seen in 3.1 below, which shows the Russian nominal inflection paradigm. This paradigm has four inflectional classes, and inflections for two number categories and six case categories. The nominative singular -*o* is predictive of all the other case forms (i.e. if you know that a given noun takes -*o* in the nominative singular you can predict its inflection in any other combination of case and number); in contrast the nominative plural -*i* is less predictive, since nouns which take that inflection show variation in inflectional marking elsewhere.

Crucially, Ackerman and Malouf (2013) observe that across natural language paradigms, while the size or e-complexity vary widely, i-complexity is consistently low. Further they show that high e-complexity paradigms tend to have low i-complexity. They conclude that i- complexity is therefore a primary measure of complexity which shapes the types of mor-

phological paradigms attested cross-linguistically.

Ackerman and Malouf (2015) further suggest that the pressure for low i-complexity shapes languages through the dynamics of language change. Specifically, during language use, low i-complexity may assist language users in solving the Paradigm Cell Filling Problem, and further, errors language users make when generalizing to unknown forms may be i-complexity-reducing. This idea is also compatible with the general hypothesis that languages evolve to maximise learnability (e.g., Christiansen and Chater 2008; Culbertson and Kirby 2016; Deacon 1997; Kirby 2002; Kirby, Cornish, et al. 2008). In this case, a learning bias against high i-complexity paradigms would drive i-complexity down over generations of learners. If i-complexity affects learning and use more than other aspects of complexity, then the former might end up being constrained across languages, while the latter may vary quite freely. That said, from this perspective the substantial variation in languages' e-complexity that Ackerman and Malouf (2013) observe is on its face surprising. We might reasonably expect that higher e-complexity also poses challenges for language learners; and the existence of languages with large morphological paradigms and numerous inflectional classes in particular is puzzling.

	SG						**PL**					
	NOM	ACC	GEN	DAT	LOC	INS	NOM	ACC	GEN	DAT	LOC	INS
noun class 1	-o	-o	-a	-u	-e	-om	-a	-a	ø	-am	-ax	-am'i
noun class 2	ø	ø	-a	-u	-e	-om	-i	-i	-ov	-am	-ax	-am'i
noun class 3	-a	-u	-i	-e	-e	-oj	-i	-i	ø	-am	-ax	-am'i
noun class 4	ø	ø	-i	-i	-i	-ju	-i	-i	-ej	-am	-ax	-am'i

Table 3.1: Russian nominal inflection paradigm (phonological transcription)[1](Baerman et al. 2010). Nouns fall into one of 4 inflection classes (rows) which show different patterns of inflection; nouns are inflected for number (SG=singular, PL=plural) and case (NOM=nominative, ACC=accusative, GEN=genitive, DA T=dative, LOC=locative, INS=instrumental).

[1]This transcription ignores the effects of an automatic rule that changes unstressed /e/ and /o/ to /i/

Here we compare how different sources of morphological complexity affect learnability of inflectional paradigms. We focus on the two types of measures described above: e-complexity as reflected in the number of inflection classes in a paradigm and the distribution of their forms, and i-complexity as reflected in the predictability of forms in a paradigm based on other parts of the paradigm. We also investigate how these interact with the number of different markers in the system, another aspect of the e-complexity of the paradigm, and different types of syncretism. Syncretism is a phenomenon in which different cells in an inflectional paradigm are realized by the same phonological form. Whether the same phonological form marks semantically related meanings or is accidental homonymy, has been suggested to affect the learning of the forms (e.g., Baerman et al. 2005; Maldonado and Culbertson 2019; Pertsova 2012). For example, in Table 1 above, -*o* is used for semantically related forms—class 1 nouns which differ in case. However, -*a* can be considered accidental homophony as it is used across different classes for different cases.

The paper proceeds as follows. We first outline more precisely how e- and i-complexity are calculated in this study. We then discuss previous work aimed at providing empirical evidence for the link between i-complexity and learning of morphological paradigms. This work has highlighted the role of predictive structure in producing novel inflections, i.e., generalization. In section 2 we then report a series of experiments using LSTM neural network and human learners testing the related hypothesis that low i-complexity provides a more general facilitatory effect on learning than e-complexity, including facilitating the retrieval of already-encountered forms early in learning. While the biases of human learners are obviously of primary interest in understanding the pressures that shape human language, we use neural networks as a convenient model of an 'ideal learner'. Testing such a learner serves to provides proof-in-principle for whether i-complexity can affect learnability and whether its

and /a/, respectively

influence is greater than other types of morphological complexity. For both human and network learners we see similar results, contrary to the hypothesis above; e-complexity generally impacts learning more than i-complexity. Finally, in section 3 we explore the relationship between the i- and e-complexity by generating a large number of random paradigms with different values of these two measures. Here we find that i-complexity and e-complexity are highly negatively correlated: as the number of distinct forms increases, the implicative structure between forms also necessarily increases. Furthermore, the range of e-complexity values is also necessarily higher than the range of i-complexity values for paradigms of the same size. These findings suggest that the observations made by Ackerman and Malouf (2013) concerning morphological paradigms may stem in part from the nature of the measures rather than pressures (e.g., inductive or usage biases) that are specially attuned to i-complexity.

3.2.1 Measuring i-complexity and e-complexity

Here we adopt methods for calculating i-complexity outlined in Ackerman and Malouf (2013). The i-complexity of inflectional paradigms is measured using the information-theoretic notion of entropy (Shannon 1963), specifically the averaged conditional entropy of forms in the paradigm. The conditional entropy of a pair of grammatical functions X,Y in the paradigm is presented in (3.1) below. Here P(x,y) indicates the joint probability of the two grammatical functions in the paradigm being realized as forms x and y, respectively; P(y|x) indicates the conditional probability of Y being realized as y, given that X is realized as x. Conditional entropy H(Y|X) quantifies the uncertainty associated with the value of Y given the value of X. For example, looking at the Russian nominal inflection paradigm above, let Y be the set of forms realizing SG.NOM, [-o, ø, -a, ø], and X be the set of forms realizing SG.DAT, [-u, -u, -e, - i]. The conditional entropy of SG.NOM given the form in SG.DAT would represent

the uncertainty associated with the form in SG.NOM, when the form realizing SG.DAT for the same lexeme is known.

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) log_2 P(y|x) \tag{3.1}$$

A paradigm's total i-complexity is the averaged conditional entropy over all pairs of grammatical functions in the paradigm, as in (3.2).[2]

$$\frac{\sum_{Y \in G} \sum_{Y \in G, X \in G} H(X|Y)}{N_G(N_G - 1)} \tag{3.2}$$

Where G is the set of grammatical functions in the paradigm and N_G is their total number.

Although Ackerman and Malouf (2013) do not explicitly suggest a measure for e-complexity, we adopt here their average cell entropy as a measure for e-complexity. The cell entropy, defined in (3.3) below, captures the number of inflection classes and the number of different variants to mark each grammatical function (e.g., combinations of number and case in the Russian nominal inflection paradigm above). Intuitively, grammatical functions that are realized with a large set of optional forms, or do not have a dominant/frequent variant, have higher cell entropy. The difference between these two measures rests in the extent to which they take into account the inter-predictability of forms across the paradigm. I-complexity is specifically defined to measure the degree to which one form can be guessed based on another form, in any other cell of the paradigm. In other words, it critically involves predicting the form of a lexeme in some grammatical function based on the form of that lexeme in a different grammatical function. By contrast, average cell entropy is only defined in terms of a single

[2]Note that this is not the only way of calculating i-complexity. For alternative formulations, see Malouf (2017) as well as Bonami and Beniamine (2016) and Sims and Parker (2016) who propose alternative formulations which are less dependent on linguist-constructed paradigms.

grammatical function, i.e., it is based on what one can predict from the form of other lexemes for that grammatical function. Average cell entropy is thus suitable for measuring what is crucially different about e-complexity as compared to i-complexity.[3] For example, Ackerman and Malouf (2013) illustrate at their claim that paradigms tend to have low i-complexity but vary in their e-complexity using the average conditional entropy and average cell entropy, respectively.

$$H(X) = -\sum_{x \in X} P(x) log_2 P(x) \tag{3.3}$$

E-complexity is measured as the averaged cell entropy over all grammatical functions in a paradigm as in (3.4).

$$\frac{\sum_{X \in G} H(X)}{N_G} \tag{3.4}$$

Where G is the set of grammatical functions in the paradigm and N_G is their total number.

[3]We further discuss the relationship between average cell entropy and another common measures of e-complexity, number of forms in the paradigm, in Section 3. In general, we prefer average cell entropy over simply counting the number of forms in the paradigm, or number of forms for a given grammatical function because the entropy-based measure also accounts for the frequency with which forms are used across a grammatical function. For example, in the Russian paradigm above, SG.GEN and SG.LOC both are expressed with two affixes, but the skewed distribution over those two affixes for SG.LOC reduces uncertainty (the appropriate affix is more likely to be -*e* than -*i*), which the entropy-based measure captures. However, it should be noted that Malouf has suggested (p.c.) that number of forms but not average cell entropy should be considered a measure of e-complexity. They argue this based on the fact that average cell entropy, like the measure of i- complexity we use, also reflects predictive relationships within the paradigm (just not across grammatical forms for a given lexeme). We would argue against this interpretation, since number of forms–an uncontroversial measure of e-complexity– can also be considered predictive in this way; it affects how well a form can be predicted based on knowledge of all the forms in the paradigm. Put another way, a paradigm with fewer forms makes any given form easier to guess.

3.2.2 Previous work investigating the effects of complexity on morphological learnability

As mentioned above, Ackerman and Malouf (2013) find evidence that while morphological paradigms differ widely in their e-complexity, the range of i-complexity values appears to be more constrained. They calculate both e- and i-complexity for inflectional paradigms in a set of 10 geographically and genetically varying languages. The e-complexity values they report (as measured by average cell entropy) ranged between 0.78 and 4.9 bits, while their i- complexity values were under 1 bit across the board.[4] A simulation analysis performed on one of the languages exhibiting high e-complexity (Chiquihuitlàn Mazatec) showed that the i- complexity of the actual paradigm was lower than the i-complexity values for random permutations of that paradigm. This suggests that the inflectional paradigms of natural languages may be organized in such a way as to minimize their i-complexity. How might this come about? One possibility is that low i-complexity facilitates solving the Paradigm Cell Filling Problem (Ackerman, James P. Blevins, et al. 2009; Ackerman and Malouf 2015), i.e., it makes it easier to determine the correct form for novel inflection. This generalization-based mechanism could lead to lower i-complexity: assuming individuals are frequently required to produce novel inflections (i.e. generate the inflectional form associated with grammatical function Y for a lexeme which they have only seen inflected for grammatical function X), and assuming they exploit predictive relationships between grammatical functions as captured by i-complexity, paradigms with low i-complexity will be relatively stable whereas paradigms with high i-complexity (i.e. where prediction from the form for function X to the form for function Y is not possible) will tend to change. Specifically, they might be expected

[4]The relationship between e-complexity and i-complexity found by Ackerman and Malouf (2013) is also reported in Cotterell, Kirov, Hulden, et al. (2019), using different measures of both e- and i-complexity (the latter based on forms drawn from corpora rather than paradigms posited by linguists, cf. Bonami and Beniamine 2016; Sims and Parker 2016)

to change in ways which reduce i-complexity since learners might actually introduce errors which reflect predictive relationships when attempting to generalise.

Seyfarth et al. (2014) tested Ackerman, James P. Blevins, et al. (2009) hypothesis that i-complexity has an effect on the ability of human learners to solve the Paradigm Cell Filling Problem. They compared the ability of human learners to predict novel inflected forms in low vs. high i-complexity input. They trained participants on an artificially constructed nominal inflectional paradigm in which nouns were marked for three grammatical numbers (singular, dual and plural) according to one of two noun classes (Table 3.2(a). In the test phase, they asked participants to generate inflected forms for a novel lexeme given that lexemes' inflected form in another grammatical number. In some trials, the required form could be predicted from the given form (predictive trials), while in others it could not be (non- predictive trials). In Table 3.2(a) for example, being prompted with a novel singular form marked with -*yez* allows the learner to predict what form the lexeme takes in the dual (-*cav*). However, knowing the form in plural is not predictive of the form in dual. They found that participants' performance differed across predictive and non-predictive trials, showing that learners were indeed able to use the predictive structure to generate a correct novel form when it was available. In a second experiment, Seyfarth et al. (2014) tested whether predictive information facilitated generalization to novel stems in a larger paradigm (Table 3.2(b)). They found that, learners made less use of predictive information in this larger paradigm: learners tended to inflect novel stems with the most frequent marker (e.g., they used the suffix -*cav* to mark dual regardless of class). Notably, while predictive relations between forms in the paradigm is captured by i-complexity, suffix frequency is captured by our measure of e-complexity. Therefore, these results suggest that e-complexity may also influence how learners generalize to novel forms.

The Seyfarth et al. (2014) study simulates a case in which language learners have to generalize

Table 3.2: Artificially constructed nominal inflection paradigms used in Seyfarth et al. (2014).

	Singular	Dual	Plural
noun class 1	-yez	-cav	-lem
noun class 2	-taf	-guk	-lem

(a) Paradigm with two noun classes (their Experiment 1).

	Singular	Dual	Plural
noun class 1	-taf	-guk	-lem
noun class 2	-yez	-cav	-lem
noun class 3	-yez	-cav	-nup

(b) Paradigm with three noun classes (their Experiment 2).

from the paradigm they have learned to inflect a novel stem for one grammatical feature based on exposure to that stem inflected for a different grammatical feature. For example, they might be required to inflect a stem for dual when they had only seen that stem inflected in the singular. They show that, in this case, learners are indeed able to use this predictive structure to predict the novel form. Johnson, Culbertson, et al. (2020) replicate these results with LSTM networks, showing that the networks are able to use the predictive relations between forms in the paradigm to generalize to novel wordforms. However, generalizing to completely novel forms is an extreme case of a much more general problem that language learners face. In addition to generalizing to completely novel forms, learners must generate (or retrieve) forms which may have been encountered but have not yet been robustly acquired. Our hypothesis is that if low i-complexity facilitates solving the Paradigm Cell Filling Problem, i.e., using familiar forms to predict new forms, it should, in principle, facilitate learning forms under low exposure as well; learners can use the same strategy they use when generalizing to completely novel stems to help generate (or retrieve) low frequency forms that are not fully memorized.

Here we test this hypothesis, comparing the effects of e- and i-complexity on the learnability of morphological paradigms. We systematically manipulate i-complexity and e- complexity,

holding other potential differences among paradigms (e.g., number of forms) constant. In Section 2, we use an artificial language learning task to train and test LSTM neural networks and human participants on four inflectional paradigms with varying values of i- and e-complexity. To test the effect of i-complexity on speed and final attainment of learning, we test how well LSTMs and human learners are able to generate forms they are trained on over the course of learning. Data from these experiments, in combination with results from Seyfarth et al. (2014), will provide evidence concerning the mechanism by which i-complexity might shape paradigms over time. Specifically, whether the pressure for low i-complexity suggested by Ackerman and Malouf (2013, 2015) comes solely from how it affects generalization to novel forms, or from a more general facilitatory effect on learning, including retrieval of encountered forms. Moreover, comparing the effects of e- and i-complexity on learning will potentially provide corroborating evidence for the hypothesis that i-complexity rather than e- complexity shapes morphological paradigms. To preview, we find that the LSTM neural networks exhibit different learning rates for paradigms with different values of i-complexity, however the effect of variations in e-complexity is larger. Results from the task with human learners reveal an effect of e-complexity but not i-complexity on learning.

3.3 Testing the effects of e- and i-complexity in human learners and LSTM neural networks

Johnson, Culbertson, et al. (2020) report a series of artificial language learning experiments with human learners and Long Short Term Memory (LSTM, Hochreiter and Schmidhuber 1997) neural networks. Learners and networks were trained on one of two nominal inflectional paradigms which were matched in e-complexity but differed in i- complexity: one with low i-complexity and one with high(er) i-complexity. They found evidence that the low

i-complexity paradigm was learned faster by LSTMs, but there was no clear effect of i-complexity for human learners. In a second series of experiments manipulating both e- and i-complexity, e-complexity was shown to better predict learnability for both LSTMs and human learners. However, in Johnson, Culbertson, et al. (2020), learning was staged, i.e., learners were first exposed to all forms in one grammatical function (singular), then forms in a second grammatical function were added (singular and plural), and finally forms in the last grammatical function were added (singular, plural, and dual). This was done to increase the chances of finding an effect of i-complexity; in low i-complexity paradigms, the dual forms could be predicted from the singular. Here, we explore more realistic, unstaged learning: presentation of forms is fully random, and learners are exposed to all forms in the paradigm from the beginning. In contrast to Johnson, Culbertson, et al. (2020), we also measure the overall accuracy of learning all inflected forms in the paradigm, rather than focusing only on learning of forms in one grammatical number. Replicating these results with unstaged learning is important, since our objective is to compare different types of complexity and their effects on learning; the learning regime should therefore be neutral in terms of enhancing or reducing the probability that learners would be affected by one measure or another. Furthermore, we take this as a starting point to investigate a wider range of differences in e- and i-complexity across paradigms, and therefore the privileged role of one specific portion of the paradigm (e.g., the singular in the staged learning design) will not hold across these more diverse paradigms.

Artificial language learning tasks allow us to create languages that differ only in the aspect we are interested in testing, while controlling for all other aspects of the language. This allows us to test the effects of i- and e-complexity on learning without confounds from other aspects of the paradigm and language such as the size of the paradigm, number of unique forms and number of words in each noun class. Another advantage of artificial languages

paradigms is that since they are small compared to natural languages, they can generally be learned to a reasonably high proficiency over the course of a single short session. While they do not reflect the full complexity of natural languages learned in natural settings, artificial language paradigms are widely used in research on language acquisition, including to investigate learning biases (e.g., Fedzechkina et al. 2012; Carla L Hudson Kam and Elissa L Newport 2009; Moreton and Pater 2012; Wonnacott and Elissa L. Newport 2005 and many others); moreover, studies using artificial learning paradigms show correspondence between lab-based learning biases and typology (e.g., see Culbertson and Elissa L. Newport 2015; Culbertson, Smolensky, et al. 2012 for reviews).

We use LSTM networks as a supplement to human learners as an additional means of testing the relative impact of i-complexity and e-complexity on paradigm learning. LSTM networks are powerful learning devices, and various recent studies show that they can be capable of extracting and using relevant linguistic information in sequence processing tasks. For example, Linzen et al. (2016) show that LSTM networks can in some cases predict long-distance subject-verb number agreement, in the presence of other potential agreement triggers (often called attractors) intervening between the subject and verb; Gulordava et al. (2018) show that LSTMs trained on four different languages can often accurately predict subject-verb agreement even when they are not trained specifically on that task; Futrell et al. (2019) show that surprisal scores of LSTMs (a measure of processing cost) paralleled preferences of human participants on grammatical judgments task differentiating word-order alternations.

Here, we use LSTMs as a convenient 'ideal learner', to provide evidence that i-complexity can in principle influence paradigm learnability for at least one learning model. This is particularly useful in circumstances where (as turns out to be the case here) human data provides little evidence of an effect of i-complexity. The LSTM models allow us to show that this is not an intrinsic limitation to the way in which we set up our learning task–we find that

i-complexity does influence learning in LSTMs trained on the same paradigms. Crucially, we can then show that, even for a class of learners sensitive to i-complexity, those effects are smaller than the effects of e-complexity. Finally, directly comparing performance of LSTMs and humans on a matched task opens up the possibility that, to the extent that they show similar patterns of performance, LSTMs could be used as a convenient tool to quickly generate predictions to be tested in further human experiments on paradigm learning. In other words, if these models reliably produce a similar pattern of results to human learners then they could potentially also be used to extrapolate to paradigms that are hard to test with human learners under controlled circumstances, e.g. learning of very large paradigms or paradigms inflecting over very large lexicons.

3.3.1 Target paradigms

We use four artificially constructed inflectional paradigms, similar in size and design to the ones used in Seyfarth et al. (2014) and Johnson, Culbertson, et al. (2020). The same paradigms were used for both neural networks and human participants. The paradigms consist of nine CVC nouns (*gob, tug, sov, kut, pid, tal, dar, ler, mip*), randomly paired with meanings for human participants (see section 3.3.3 below). The nouns were randomly allocated to three classes (for each run of the network, or each human participant), and each class was inflected for three numbers: singular, dual and plural. Inflectional markers were seven VC monosyllabic suffixes (*-op, -oc, -um, -ib, -el, -ek, -at*). These inflectional markers were randomly allocated to cells in each paradigm, such that the four paradigms were always structured as in Table 3.3 below but with a different mapping of affixes to cells for each human participant.

As summarized in Table 3.3, the paradigms differ either in i-complexity or e-complexity,

		e-complexity	
		Low (1.141 bits)	**High** (1.363 bits)
i-complexity	**Low** (0.222 bits)	low-i/low-e	low-i/high-e_{within} low-i/high-e_{across}
	High (0.444 bits)	high-i/low-e	

Table 3.3: Four target paradigms differing either in i-complexity or e-complexity values. The low i-complexity, low e-complexity (low-i/low-e) and high i-complexity, low e-complexity (high-i/low-e) paradigms differ in i-complexity only. The two remaining low-i/high-e paradigms have low i-complexity but have higher e-complexity; these paradigms also differ in the type of syncretism pattern (within class or across class).

holding the other constant. We also hold constant all other aspects of the paradigms: the paradigms are matched in terms of number of distinct affixes and number of inflectional classes, and they feature the same three-way number distinction. The low i-complexity, low e- complexity (low-i/low-e) and high i-complexity, low e-complexity (high-i/low-e) paradigms differ in their i-complexity (0.222 vs. 0.444 bits) while their e-complexity is kept constant (1.141 bits). The key difference between the two paradigms is that in the low-i/low-e paradigm, knowing the singular affix of a word (e.g., -*op* in Table 3.4a), predicts the dual affix (e.g., -*um*). This is not the case in the high-i/low-e paradigm (in Table 3.4b the singular -*op* does not uniquely determine the form of the dual). The remaining two paradigms (Table 3.4c, 3.4d) both have low i- complexity (0.222 bits) but higher e-complexity (1.363 bits). In general, higher e-complexity here means having distinct dual forms for each class, which results in a higher uncertainty across forms relative to the low e-complexity paradigms. I-complexity is kept constant and low in these two paradigms since both the plural and dual forms are predictive of each other as well as the forms in singular. However, increasing e-complexity while keeping the number of markers constant requires *syncretism* in the paradigm; a single affix is used to mark different grammatical functions. In order to additionally explore

	Singular	Dual	Plural
noun class 1	-op	-um	-ib
noun class 2	-at	-oc	-el
noun class 3	-op	-um	-ek

(a) low-i/low-e

	Singular	Dual	Plural
noun class 1	-op	-um	-ib
noun class 2	-at	-um	-el
noun class 3	-op	-oc	-ek

(b) high-i/low-e

	Singular	Dual	Plural
noun class 1	-op	-um	-op
noun class 2	-at	-ib	-el
noun class 3	-op	-oc	-ek

(c) low-i/high-e_{within}

	Singular	Dual	Plural
noun class 1	-op	-um	-el
noun class 2	-at	-ib	-op
noun class 3	-op	-oc	-ek

(d) low-i/high-e_{across}

Table 3.4: Example paradigms for each type tested. See Table 3.3 for high-level descriptions of each type. Colored cells highlight distinct paradigm structures: in low-i/low-e (a), singular *-op* predicts dual *-um*; in high-i/low-e (b), singular does not predict dual; in both low-i/high-e paradigms (c,d), the singular form which occurs most frequently is reused for plural elsewhere in the paradigm (syncretism)—either in one of the classes with that form in the singular (c low-i/high-e_{within}), or in a different class (d low-i/high-e_{across}).

how syncretism affects learning, here we generated two different syncretism patterns: within class syncretism (low-i/high-e_{within}) and across class syncretism (low-i/high-e_{across}). In both low-i/high-e paradigms, the singular form is the same for classes 1 and 3 (e.g., -op in the example paradigm in Table 3.4c, 3.4d). In the low-i/high-e_{within} the syncretic form is reused as a plural in class 1 (Table 3.4c). In the low-i/high-e_{across} the syncretic form is reused as a plural marker for class 2 (Table 3.4d)–crucially, not one of the classes which use this form in the singular. Previous work on morphological paradigms suggests that this difference in syncretism type could affect learning in human learners (e.g., Baerman et al. 2005; Maldonado and Culbertson 2019; Pertsova 2012), therefore we test both paradigm types.

Note that we do not include a paradigm with high i-complexity *and* high e-complexity. This is not actually possible: there is no way to distribute markers such that both measures of complexity are high without changing the number of markers in the paradigm. We discuss this further below.

As mentioned above, in Johnson, Culbertson, et al. (2020), exposure to forms from a paradigm was *staged*: input initially contained only singular forms, then singular and plural forms, then singular, plural, and dual forms. This was designed to highlight the implicative structure of low i-complexity paradigms. However, it is also rather unrealistic in that exposure in natural language is unlikely to be staged in this way, or at least not so rigidly staged. Here, we expose learners to forms drawn at random from the entire paradigm. Therefore, we test whether having low vs. high values of i- or e-complexity is beneficial when learners have not always learned predictive forms first. We compared speed and accuracy of learning all forms in the language across all four conditions.

3.3.2 Experiment 1: LSTM neural networks

Neural networks are computational models which approximate a function linking the network's input with its desired output. The model consists of several layers of nodes interconnected by associative weights. Given a dataset of input-output pairs the model tries to learn the optimal setting of these weights to correctly transform an input into its corresponding output. Updating the weights to better approximate the input-output function is done by searching for weights that minimize the *loss function* of the network, which measures how close the network's output is to the true output. Different algorithms are used for this search. A common algorithm is *(stochastic) gradient descent*. Intuitively, the network generates an output through a forward pass from the input layer to the output layer, after which the loss function calculates the difference between the predicted and the target values. Then in a backward pass the loss function is used to compute an error gradient with respect to each weight and the network's weights are updated in the direction of the greatest descent so as to reduce this error.

Recurrent neural networks (RNNs) overcome a limitation of simple neural networks fundamental to language tasks; simple neural networks are not sensitive to the 'context' of the current input or, in other words, how previous inputs affect the correct output for the current input. RNNs overcome this limitation by having 'short term memory' through looping back the output or hidden layer activations previously produced for earlier inputs (Jeffrey L. Elman 1990; Jeffrey L. Elman 1991; Jordan 1997). This allows the network to adjust the output for the current input according to previous inputs. The extent to which previous states of the network affect the current state is also determined by weights updated through the backpropagation process.

Long Short Term Memory (LSTM) networks are an extension of recurrent neural networks introduced by Hochreiter and Schmidhuber (1997) in order to improve learning of longer temporal dependencies. Practically, LSTMs add an element of 'long term memory' to networks by allowing the network to control the influence of current and previous inputs during the process of activation propagation, using 'weighted gates' in the networks. Like activation weights, these gates are optimized during training to determine what information is stored or passed along and therefore allowed to influence subsequent inputs. This allows LSTMs to make better use of sequential information, including learning non-adjacent dependencies.

LSTMs therefore offer a powerful but convenient general-purpose learning mechanism for language based tasks. Here we use LSTMs to process relatively short sequences: networks are presented with stems and grammatical features and produce an inflectional affix, and we train models on the target paradigms which differ in either their i-complexity or e-complexity.

Network structure

We trained and tested LSTM networks using the Keras package in Python (Chollet et al. 2015). In this task, the model gets as input a sequence containing the noun's stem and an extra character indicating the grammatical number of the object (1 for singular, 2 for dual and 3 for plural). For example, the string *mip3* indicates the noun with the stem *mip* in plural. The model's task is to output the correct affix for this wordform, according to the paradigm it is trained on. The network has 7 output units, one for each of the 7 affixes in the target paradigms. Input stem+number sequences are encoded as one-hot vectors. i.e., every character used in the language is represented as a vector of zeroes (with length equal to the total set of characters, 27) with '1' in a different index uniquely identifying it. We trained the model with a range of embedding vectors dimensionalities for the input layer and LSTM hidden layer dimensionalities (from 5-dimensional embedding vectors and 5-unit layer (542 parameters) to 50 (14,657 parameters), with increases of 5 units). The state of the LSTM at the end of the input string is fed into a 'softmax' function to produce a one-hot encoding representing the output affix for this stem+number input (i.e. the network's task it to learn a 7-way categorical classification of the input sequences). The network was optimized using Stochastic Gradient Descent (SGD) with learning rate of 0.1, batch size of 32, and no dropout.[5] Initial weights were randomly generated, according to a 'glorot_uniform' function (sampling from a uniform distribution in the range of $[-x, +x]$, where x is a function of the size of the network).

For each paradigm and set of hyperparameters, 50 runs were produced. In each run, the lexical items were randomly assigned to noun classes and the model was trained and tested

[5]In addition to the various network sizes reported in the main paper, we also ran variants of the model with a range of learning rates, using both SGD and Adam (Kingma and Ba 2014) optimizers. Detailed results are presented in the Appendix, note that the overall conclusions discussed in the main text remain unchanged across these variants.

on input-output pairs across 900 epochs . In each epoch, the network is trained and tested on all 27 wordforms in the language (9 stems marked for singular, dual and plural). The test set in this task is identical to the training set—we are not testing the capacity of the network to generalize, but rather the overall accuracy and speed with which it learns the mapping from stem+number input to the appropriate affix output.[6]

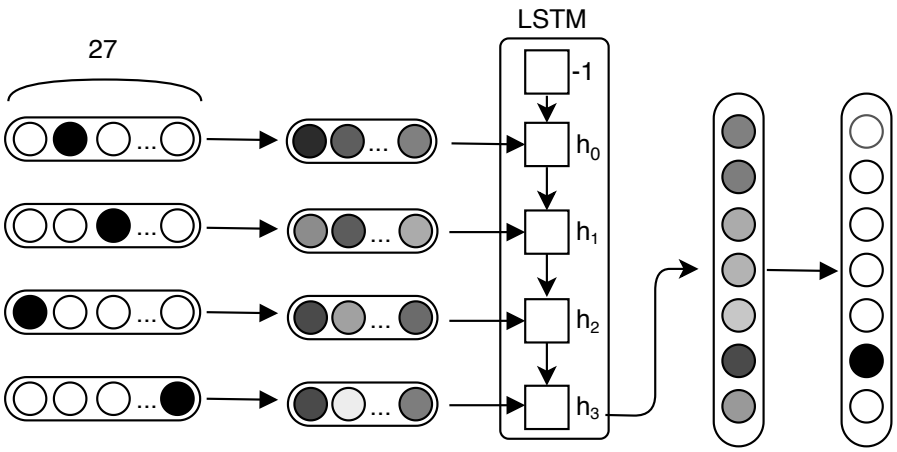


Figure 3.1: A diagram of the recurrent neural network: the input layer receives a string of four characters (stem + grammatical number), each coded as a one-hot vector of the length of the different characters used in the language (27). The input vectors are embedded and the embeddings are transferred to a hidden layer with 5-50 LSTM units. Output from the LSTM units (h_3) is then transferred to an output layer with seven options, representing the seven suffixes in the language. Using a softmax function, the output is converted to a one-hot vector, representing the suffix the network selected for this input.

Results

We measured the average accuracy of the networks in producing the correct affix for all wordforms in the target paradigm over epochs (averaged over 50 runs for each combination of target paradigm and network size). For simplicity, we first collapse the two low-i/high-e

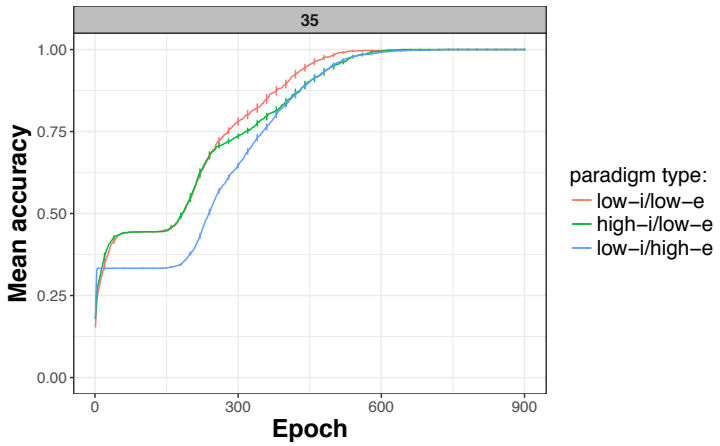
[6] As discussed above, this task differs from that used in Seyfarth et al. (2014), who focus on generalizing to unknown forms.

paradigms in these graphs, and deal with the effect of syncretism separately below. Figure 3.2 presents network learning trajectories for these three paradigm types. The same trend is seen across different network sizes. While 900 epochs is sufficient for all paradigms to be learned perfectly, even for the smallest networks, the low-i/low-e paradigm type is learned most rapidly. Networks trained on the high-i/low-e paradigm type show a similar but slightly slower learning trajectory. Networks trained on the low-i/high-e paradigm types show the slowest learning, with accuracy increasing markedly later in training than the other paradigms.

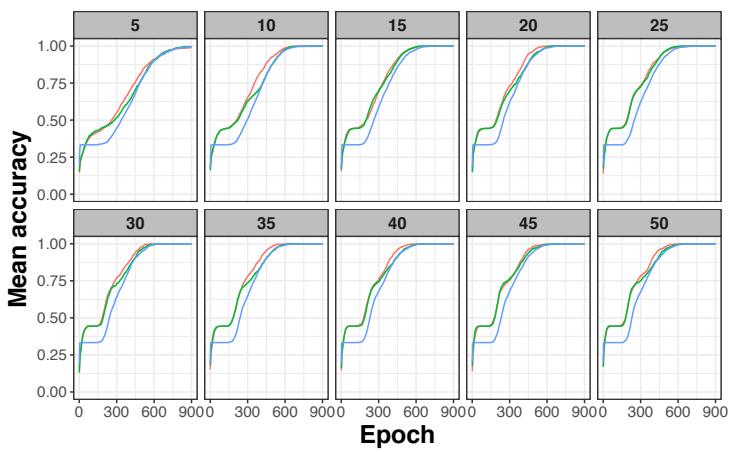
Since we are interested in the effect of i- and e- complexity on the difficulty of learning the paradigm, rather than whether the language is eventually learnable or not (all of our paradigms were eventually learned with 100% accuracy given sufficient training), we compare the *summed accuracy* (i.e. the sum of the epoch-by-epoch accuracies) of the networks trained on the different languages. The summed accuracy reflects both the speed of learning the language and the accuracy throughout learning. For example, in the results shown in Figure 3.2, where all networks eventually reach ceiling, networks which learn more rapidly will have a higher summed accuracy reflecting the faster pick-up in accuracy over epochs. Other measures of learning speed are possible, e.g. the mean number of epochs to reach 100% accuracy; we prefer mean summed accuracy because it relates more obviously to the different shapes of curve we see in Figure 3.2, and is still interpretable for network parameterisations that do not result in convergence to 100% accuracy.

Figure 3.3 shows the summed accuracy of the networks trained on each paradigm type

[7]We looked at the errors made by the LSTMs at epochs 1-150 (when the neural networks show a plateau in learning). At this stage in learning, the networks use only two out of the seven possible affixes as an output. This likely reflects a local minimum in the loss function, meaning that the LSTM 'found' a partial solution that maximizes its output accuracy. Each input string is classified with one of those two affixes solely according to the number indicating the grammatical number at the end of the input string so that all singulars take one affix (one of the affixes that mark singular), and all dual and plural inputs are marked with another affix (one of the affixes that mark either dual or plural). After around 150 epochs, the networks start using additional affixes, which is then reflected by a jump in performance.



(a)



(b)

Figure 3.2: Network learning trajectories.[7](a) results for one network size (35 cells), with error bars indicating standard error every 10 epochs, (b) results for all network sizes tested (facet titles give network size in number of cells). Networks trained on low-i/low-e and high-i/low-e paradigm types show similar learning trajectories, while networks trained on low-i/high-e paradigms show lower accuracy levels. Results from models with further learning rates for both SGD and Adam optimizers show similar patterns for most cases, and we never see the opposite trend of lower accuracies for the high-i/low-e condition (see Appendix for detailed results).

across different network sizes. To determine whether these differences between network learning trajectories are significant, we ran a linear mixed-effect regression model[8] predicting

[8]All models reported here were run using the lme4 (Bates et al. 2014) and lmerTest (Kuznetsova et al. 2017) packages in R.

the summed accuracy of the network across all epochs based on fixed effects of paradigm type (low-i/low-e, high-i/low-e, low-i/high-e), size of the network, and their interaction. In addition to these fixed effects, we also included random intercepts for each run of a network. Network size was mean centred. Paradigm type was Helmert-coded to test our predictions about the relative levels of accuracy across paradigms. Based on results from Johnson, Culbertson, et al. (2020) we predict low-i/low-e to be easiest, therefore this was set as the baseline. The model compares the baseline to the next level, high-i/low-e, then the mean of these two levels is compared to the third level, low-i/high-e. The first contrast therefore tests the effect of i-complexity and the second tests the effect of e-complexity. The model revealed a significant effect of network size on summed accuracy ($\beta = 1.63, sd = 0.049, t = 32.83, p < 0.001$), suggesting that larger networks learn the languages faster. Critically, the model also revealed a significant effect of both i-complexity ($\beta = -4.48, sd = 0.9, t = -4.68, p < 0.001$) and e-complexity ($\beta = -10.61, sd = 0.52, t = -20.23, p < 0.001$) on summed accuracy. These results suggest that measures of paradigm complexity based on implicative structure (i-complexity) and on number and distribution of forms (e-complexity) both impact ease of learning in LSTM neural networks. Note that while both effects are significant, the estimated effect size for the effect of e-complexity is larger than the estimate effect of i-complexity, suggesting the e-complexity manipulation had a larger effect than our i-complexity manipulation; this difference in effect sizes can be seen in the timecourses in Figure 3.2 and in Figure 3.3.

Type of Syncretism

Recall that we included two types of low-i/high-e paradigms: one in which syncretism was within class, and one where it was across class (see Table 3.4). In general cross-class syncretism can affect both i-complexity and e-complexity, but for our paradigms neither i-

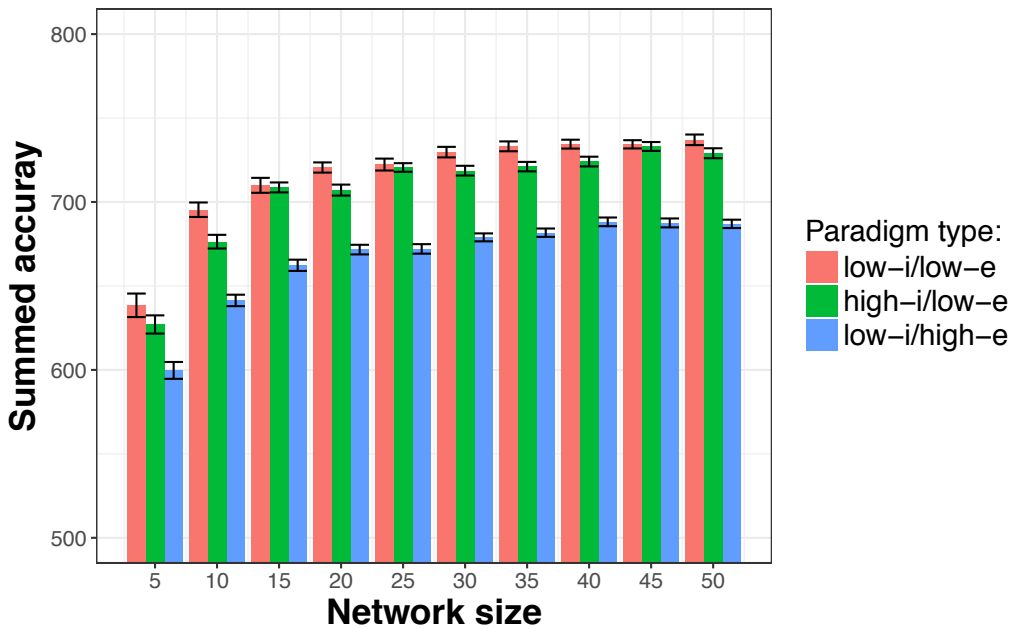
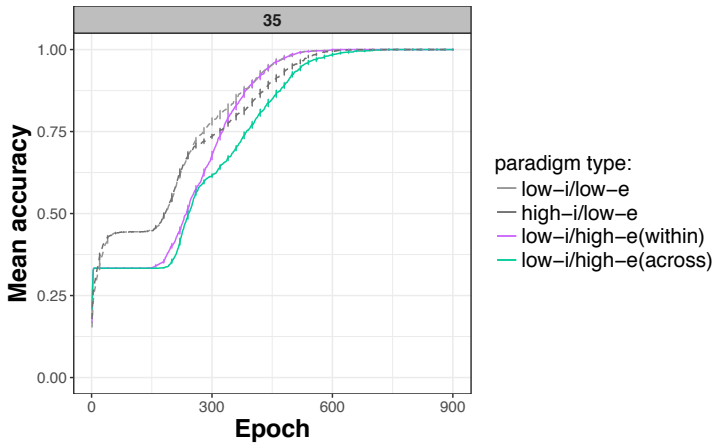


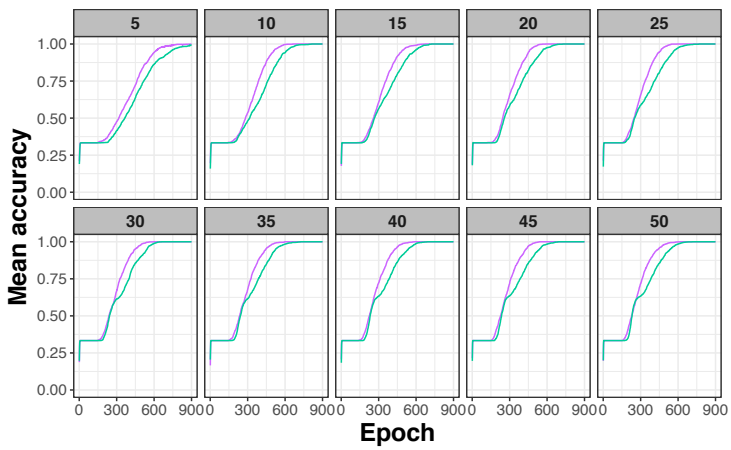
Figure 3.3: Summed accuracy over the 900 epochs of the networks trained on each of the three paradigm types across different sizes of the network. Note that the two low-i/high-e paradigms are collapsed here.

complexity nor e-complexity distinguish between syncretism types; the two paradigm types have the same values for both measures. Figure 3.4 shows network learning trajectories with these two paradigm types plotted separately. Across different network sizes, the paradigm type with cross-class syncretism appears to be learned slower, in line with previous work (e.g.,Maldonado and Culbertson 2019; Pertsova 2012).

Summed accuracies of networks trained on low-i/high-e_{within} and low-i/high-e_{across} paradigms (averaged over the 50 runs of the model) across different network sizes are presented in Figure 3.5. We ran a linear mixed-effect regression model predicting summed accuracy by paradigm type (within-class syncretism vs. across-class syncretism), network size and their interaction. In addition to these fixed effects, random intercepts for each run of a network. Paradigm type was dummy coded, with within-class syncretism coded as the reference group. Network size was mean centred. The model revealed a significant effect for the network size, increasing



(a)



(b)

Figure 3.4: Network learning trajectories with low-i/high-e_{within} and low-i/high-e_{across} paradigms plotted separately. Trajectories for networks trained on low-i/low-e and high-i/low-e paradigms presented in grey (dashed lines) for comparison. (a) results for one network size (35 cells), with error bars indicating standard error every 10 epochs. (b) results for all network sizes tested (facet titles give network size in number of cells). Networks trained on paradigms with cross- class syncretism show slower learning.

the learning accuracy for larger neural networks ($\beta = 1.45, sd = 0.09, t = 15.9, p < 0.001$). Critically, the model also revealed a significant effect of paradigm type ($\beta = -34.37, sd = 1.84, t = -18.62, p < 0.001$), suggesting that paradigms with across-class syncretism are learned slower by the neural networks.

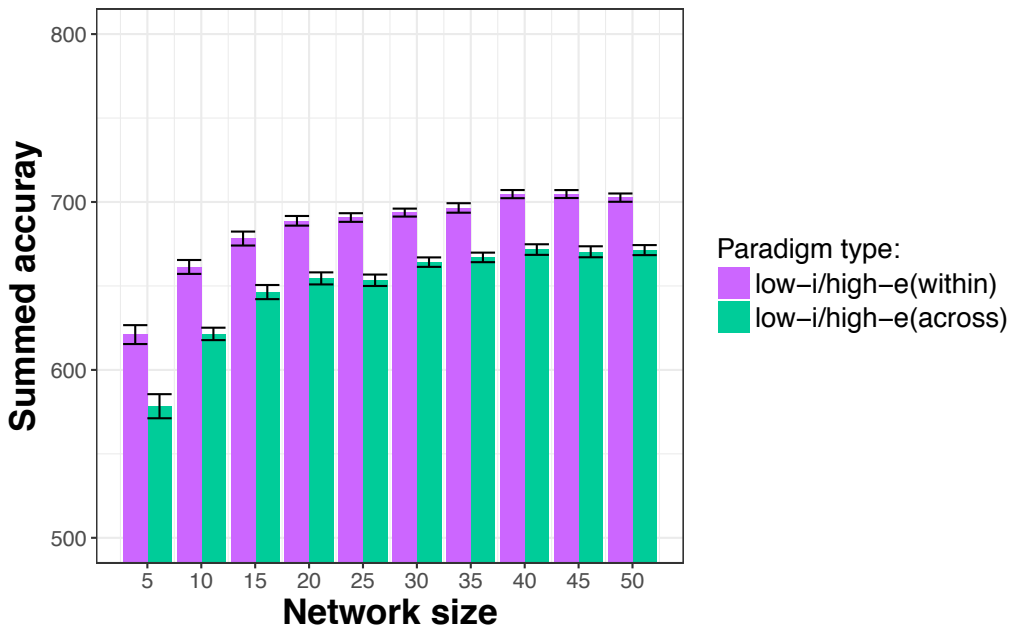


Figure 3.5: Summed accuracy over the 900 epochs of networks trained on low-i/high-e_{within} and low-i/high-e_{across} paradigms across different network sizes. Across all network sizes the paradigm type with across-class syncretism is learned slower.

Since the type of syncretism was found to affect learning, we conducted an additional analysis to determine whether the effect of e-complexity was entirely driven by the low-i/high-e_{across}, or whether this effect is found regardless of syncretism type. We ran a linear mixed-effect regression model predicting summed accuracy by paradigm type and network size (mean centred), with random effects as specified for previous models. Paradigm type was dummy coded with low-i/low-e as the reference group. The model revealed a significant effect of network size ($\beta = 1.61, sd = 0.09, t = 17.25, p < 0.001$). In addition, the model revealed a significant difference between low-i/low-e and both low-i/high-e paradigm types (low-i/high-e_{within}: $\beta = -31.3, sd = 1.89, t = -16.52, p < 0.001$, low-i/high-$e_{across}$: $\beta = -65.67, sd = 1.89, t = -34.67, p < 0.001$). This confirms the generality of the effect of e-complexity on learning; regardless of the type of syncretism, paradigms with high e-complexity are learned more slowly than languages with low e-complexity, even when all other aspects of

the paradigm (i-complexity, but also number of inflections, number of inflectional classes, etc.) are held constant. As before, there was also a significant difference between low-i/low-e and high-i/low-e ($\beta = -8.96, sd = 1.89, t = -4.73, p < 0.001$).

To summarize, here we trained LSTM neural networks on one of four nominal inflectional paradigms which differed in either i-complexity or e-complexity. The results of our simulation experiments showed that both measures of complexity affect learning in these networks, with more complex paradigms being learned more slowly. We also found that type of syncretism mattered: networks more readily learned syncretic forms which targeted cells within a class rather than across class. These effects were not necessarily all of equal strength: effects of i-complexity were weaker than the effects of e-complexity and syncretism type. The effect size of e-complexity on the network's accuracy was four times larger than the effect of i-complexity (estimated β values of -31.3 in the case of within-class syncretism and -65.67 in the case of across-class syncretism vs. -8.96 for the effect of increased i-complexity). In sum, our neural network simulations show that, in principle, i-complexity can affect learning morphological paradigms. This complement earlier results for human learners and LSTMs (Johnson, Culbertson, et al. 2020; Seyfarth et al. 2014) showing that low i-complexity facilitates generalisation to novel forms. Importantly however, our results also provide evidence that e- complexity has a stronger effect on learning. In the next section, we turn to human learners. Johnson, Culbertson, et al. (2020) found that i-complexity only weakly affected human learning, even in a staged paradigm intended to maximise the effects of i-complexity. Here we will compare the effects of i- and e-complexity to see whether indeed e-complexity plays a stronger role in determining ease of learning for humans when learning is not staged.

3.3.3 Experiment 2: human learners

Materials

The same artificially constructed paradigms described in Table 3.4 were used to train and test human participants. Participants were exposed to the word forms in the language together with meanings. Stems referred to a set of simple objects: lemon, cow, tomato, bicycle, horse, clock, pigeon, mug and pear. Visual stimuli were identical to those used in Johnson, Culbertson, et al. (2020). Singular nouns corresponded to a single object, dual corresponded to two objects, and plural ranged from 3-12 objects (selected randomly). See Figure 3.6 for an example plural trial. Objects in the language were divided into the three noun classes so that every noun class had one animate object (cow/pigeon/horse), one edible object (tomato/lemon/pear) and one other (clock/bicycle/mug). This was done to ensure that noun class membership could not be determined based solely on semantic features. All stems and markers were randomly assigned to meanings for each participant.

Participants

144 self-reported native English speakers participants were recruited via Amazon's Mechanical Turk crowd-sourcing platform. They were compensated $6 for their participation and the experiment lasted 53 minutes on average (min = 19, max = 166, mode = 41). We recruited participants who possessed an Mturk qualification indicating that they were based in the US. Participants were allocated randomly to each of the four paradigms. We excluded from the final dataset 22 participants who did not complete the experiment[9], thus the final dataset consisted of 120 participants: low-i/low-e (29); high-i/low-e (31); low-i/high-e_{within}

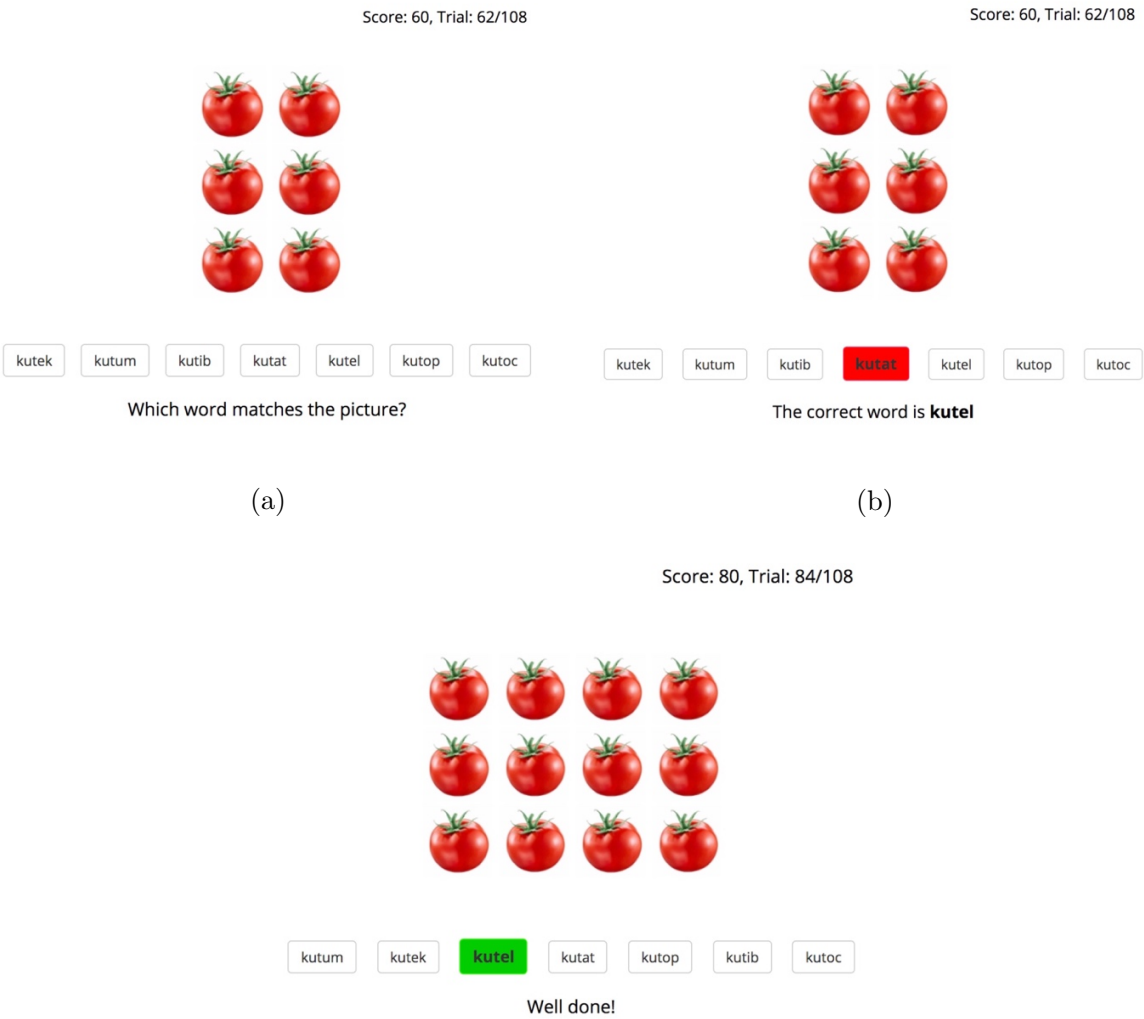


Figure 3.6: Example plural trial. (a) a picture is presented and participants are asked to choose the correct label from a set of options. (b), (c) participants receive feedback on their answer, including the correct label. (b) negative feedback following trial shown in (a), (c) positive feedback following plural trial with a different number of objects.

(28); low- i/high-e_{across} (31).

Procedure

Participants learned the language via trial and error. On each trial, a picture (featuring 1-12 instances of a single object) was presented on the screen together with a set of possible labels, as in Figure 3.6. Participants were asked to choose the correct label after which they received feedback on their answer. If their answer was incorrect, they were presented with the correct form. The set of possible labels consisted of all combinations of the correct stem with all the suffixes in the paradigm. The task was divided into 3 identical blocks of 108 trials each: in every block, participants were exposed to all stems inflected in each of the three grammatical numbers (27 wordforms), 4 times each. The order of trials was randomized in each block. Participants were allowed a self-paced break between blocks; they were presented with a screen announcing the end of the block and were asked to click on 'continue' to complete the next block of trials. Participants' answers on each trial were recorded and their overall accuracy was measured to test the effects of i-complexity and e-complexity on paradigm learnability.

Results

Figure 3.7 shows learning trajectories for each paradigm type, here with low-i/high-e paradigm types (which differed in syncretism type) collapsed. Participants' learning trajectories are non-linear but less complex than the learning curves of the LSTMs and can be described using quadratic polynomial curves (as in Figure 3.7). Therefore, we used logistic growth curve analysis (Mirman 2017) to analyse the effect of i-complexity and e-complexity on learning

[9]Participants who did not complete the experiment and who contacted us were paid according to the proportion of trials they completed.

over trials. The model predicted accuracy by paradigm type and trial number. In addition to these fixed effects, the model also included by-participant intercepts and random slopes for trial number. Paradigm type was Helmert-coded as in Experiment 1. Learning curves (accuracy over trials) were modelled with second-order orthogonal polynomials. The model revealed no significant effect of i-complexity ($\beta = 0.2, sd = 0.15, z = 1.29, p = 0.19$), but a significant effect of e-complexity ($\beta = -0.16, sd = 0.07, z = -2.18, p = 0.028$):participants trained on one of two low e-complexity paradigms learned better than participants trained on a high e-complexity paradigm. There was also a significant effect of trial in both the linear ($\beta = 9.9, sd = 0.87, z = 11.3, p < 0.001$) and quadratic ($\beta = -2.23, sd = 0.43, z = -5.16, p < 0.001$) terms, indicating that across trials, overall accuracy increased, but curves became less steep over time. These results provide clear evidence of the effect of e-complexity on human learning of inflectional paradigms. However, our results fail to show any effect of i-complexity. The data are noisy, but the numerical trend is in fact in the wrong direction–the high-i/low-e paradigm is learned numerically better than the low-i/low-e paradigm.

One plausible strategy, which would be consistent with the results showing an effect of e-complexity and no evidence for an effect of i-complexity, is simply to choose the most frequent form for each grammatical number, ignoring class membership for each stem. This strategy would result in higher accuracy in the low e-complexity conditions (where there is a frequent form for both the singular and the dual, see Table 3.4) but would yield lower accuracy in the high e-complexity conditions (where there is a frequent form in singular only). However, a closer look at our participants' responses, and the rates with which they chose the frequent form for each grammatical number, show that this is probably not the case; participants (as a group) do not choose the frequent form for a specific number more than its actual probability with which is appears (66% of the trials with this grammatical number). Participants in the low-i/low-e condition on average chose the frequent form of

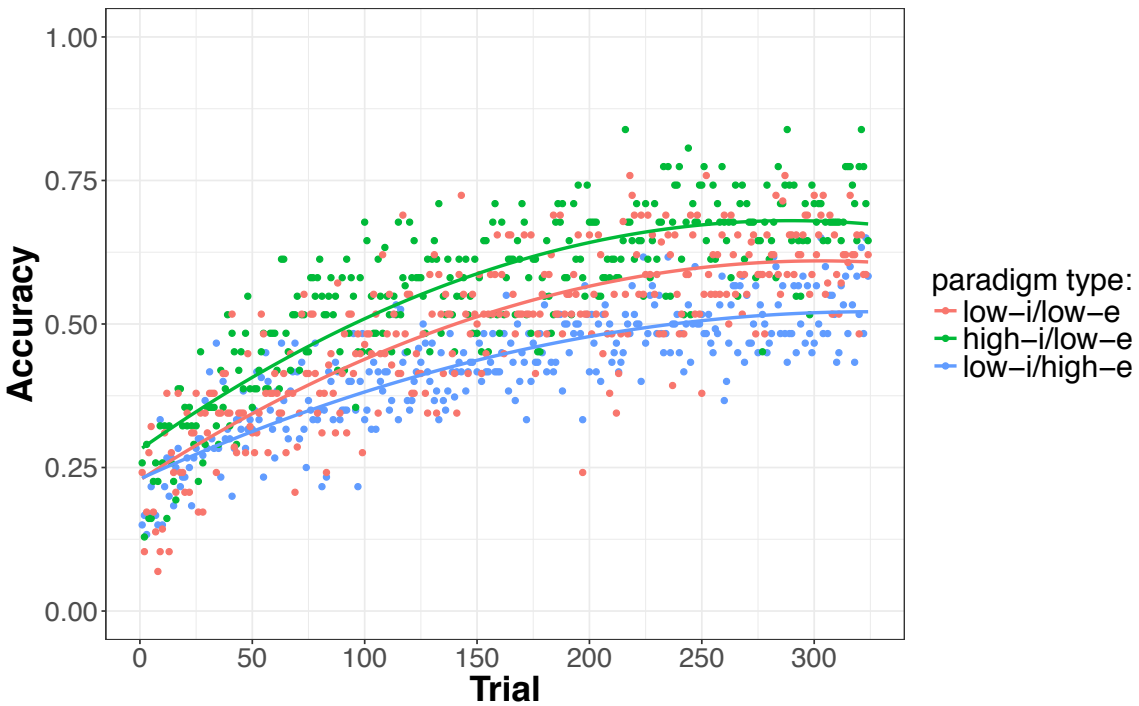


Figure 3.7: Mean accuracy by trial for each of the three paradigm types (collapsing the two low- i/high-e paradigms). Points indicate the average accuracy across participants for each trial. Lines show quadratic polynomial curves predicting accuracy by trial number for each paradigm type. Learning is worst for the low-i/high-e and best for the high-i/low-e paradigms.

a grammatical number in 64.9% of the relevant trials, and participants in the high-i/low-e condition chose the frequent form of a grammatical number in 66.5% of the relevant trials. These results suggest that participants are probability matching (e.g., Carla L. Hudson Kam and Elissa L. Newport 2005; Carla L Hudson Kam and Elissa L Newport 2009); participants match the probability of the form in their responses to its actual probability in the language rather than simply choosing the most frequent form for each grammatical number. Therefore, there is an advantage to the skewed distribution of forms in low e-complexity paradigms that facilitates learning the paradigm even if participants do not simply select the most frequent form.

Type of syncretism

As with the LSTMs, we further tested whether there was a difference in learning for the two paradigms differing in syncretism type. We ran a separate logistic growth curve model predicting accuracy by paradigm type (within-class syncretism vs. across-class syncretism, sum coded) and trial number, with by-participant intercepts and random slopes for trial number. Here as well, learning curves (accuracy over trials) were modelled with second-order orthogonal polynomials. The model revealed no significant effect of syncretism type ($\beta = -0.019, sd = 0.15, z = -0.127, p = 0.89$). As before, the model revealed a significant effect of trial in both the linear ($\beta = 8.06, sd = 1.19, z = 6.9, p < 0.001$) and quadratic ($\beta = 8.06, sd = 1.19, z = 6.9, p < 0.001$) terms, indicating that across trials, overall accuracy increased, but curves became less steep over time. The results do not provide any evidence for differences in learnability of morphological paradigms with across-class as compared to within-class syncretism in human learners. There is therefore no reason to suspect that the effect found above of e-complexity in human learners is driven by differences in learnability across types of syncretism.

3.4 Exploring the relationship between i- and e-complexity with random paradigms

Results from simulations with LSTM neural networks and behavioural experiments with human learners both suggest that e-complexity has a robust effect on learning of inflectional paradigms. By contrast, the effect of i-complexity was present but weaker in neural networks and absent in human learners. This suggests that i-complexity is not the primary determinant of learnability–e-complexity, at least how we have measured it here, has a much larger impact

on how well learners are able to generate (or retrieve) forms they have been exposed to. It may be that the beneficial effects of low i-complexity largely derive from its facilitating effect on generalisation (as suggested by Ackerman and Malouf (2015)). Ackerman and Malouf (2013)'s Low I-complexity Conjecture for natural languages is based on the observation that, across a sample of natural languages, a relatively wide range of e-complexity values was found, but the range of i-complexity values was much more narrow. From this Ackerman and Malouf (2013) concluded that e-complexity in natural morphological paradigms is relatively free to vary and can be high as long as i-complexity stays low. However, as we have already mentioned, these two measures are not independent of one another: it was not possible for us to construct a paradigm with both high e-complexity and high i-complexity (while keeping the number of forms constant). In this section we explore the relationship between i- and e-complexity by looking at their values across 1000 randomly generated paradigms. To preview, we find an inverse correlation between i- and e-complexity which is in line with the pattern Ackerman and Malouf (2013) observe. This suggests that the Low I- complexity Conjecture is not necessarily a result of language change, i.e., it may not be driven purely from usage errors or learnability pressure. We also test the learnability of this set of 1000 paradigms with LSTM neural networks to show how these two measures relate to learning across a wider range of paradigms than we covered in Experiments 1-2.

3.4.1 Generating random paradigms

We generated 1000 random inflectional paradigms expressing the same three grammatical numbers (singular, dual and plural) across three noun classes, as in the paradigms tested above. The paradigms were generated by randomly assigning affixes to the nine cells with replacement, i.e., allowing affixes to repeat. Therefore, paradigms also vary randomly in number of unique affixes. Generated paradigms had between three and eight affixes, with

most paradigms (42%) including six unique affixes. For each randomly generated paradigm, we calculated i- and e-complexity. I-complexity varied between 0 and 0.667 bits with a mean value of 0.201 bits. E-complexity varied between 0.528 and 1.585 bits with a mean value of 1.36 bits.

3.4.2 Quantifying the relationship between i- and e-complexity in random paradigms

We first explored the relationship between these three dimensions of variation (i-complexity, e-complexity, number of distinct affixes) in the 1000 randomly generated paradigms. Figure 3.8 shows the distribution of i-complexity and e-complexity values across paradigms, with average number of markers indicated by color. As suggested by the figure, i-complexity is strongly negatively correlated with e-complexity ($r = -0.92, t(998) = -73.8, p < 0.001$). In other words, paradigms with high i-complexity tend to have low e-complexity, and vice versa. To explore the relationship between these complexity measures and the number of the unique affixes in the paradigm, we ran additional correlation tests. While e-complexity is positively correlated with the number of markers in the paradigm, ($r = 0.44, t(998) = 15.62, p < 0.001$), i-complexity is negatively correlated with it ($r = -0.38, t(998) = -13.1, p < 0.001$): as the number of distinct forms increases, the implicative structure between forms increases. For example, if every cell in the paradigm is expressed by a unique form, then each form will perfectly predict every other form.

Since both i-complexity and e-complexity correlate with the number of markers in the paradigm, we further analysed the subset of random paradigms with the most frequently generated number of markers (six). We tested the relationship between i-complexity and e-complexity for these paradigms (423 paradigms), again confirming the negative correla-

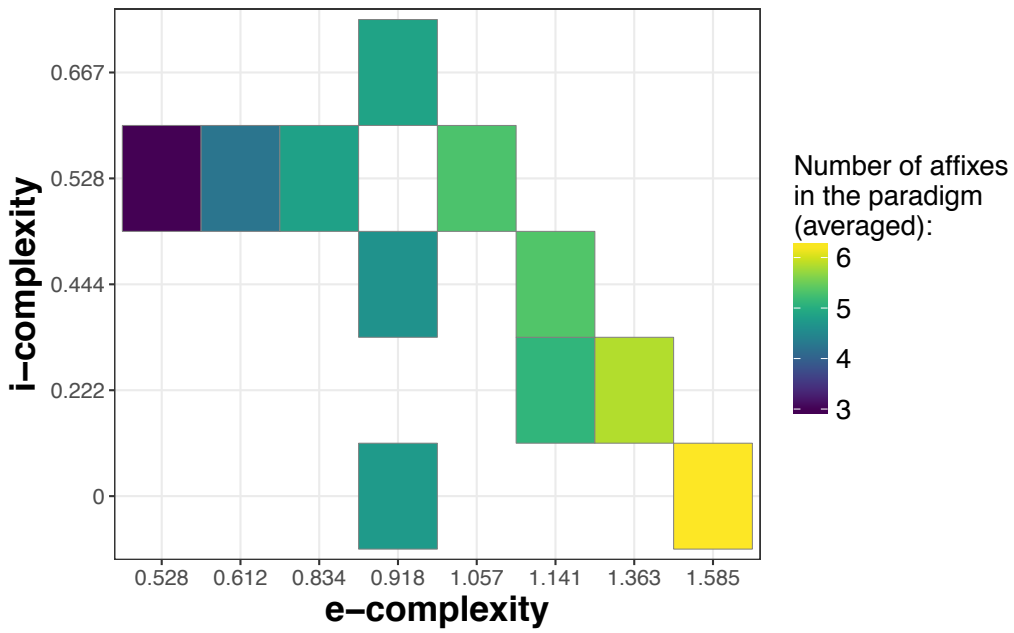


Figure 3.8: Distribution of randomly generated paradigms in terms of i- and e-complexity. Color represents the average number of markers for paradigms with specific i- and e-complexity values. No paradigms have high i-complexity *and* high e-complexity. Paradigms with high i-complexity and low e-complexity have on average fewer markers while paradigms with low i- complexity and high e-complexity have more.

tion ($r = -0.94, t(421) = -59.24, p < 0.001$). Table 3.5 presents two randomly-generated example paradigms with six markers which illustrates how the negative correlation between i-complexity and e-complexity arises from the organization of markers in the paradigm, even when the number of markers in the paradigm is held constant. Paradigms in which a grammatical function is marked with the same marker across inflection classes tend to have lower e-complexity (there is a more frequent form marking this grammatical function) and higher i-complexity (forms in this grammatical function are less likely to predict other forms in the paradigm).

The strong negative correlation between i-complexity and e-complexity has clear implications for how Ackerman and Malouf (2013) findings should be interpreted. They show that across a sample of morphological paradigms in ten languages, e-complexity reaches relatively high

	Singular	Dual	Plural
noun class 1	6	5	6
noun class 2	8	1	3
noun class 3	5	7	7

(a)

	Singular	Dual	Plural
noun class 1	2	6	8
noun class 2	4	0	8
noun class 3	1	6	8

(b)

Table 3.5: Two example paradigms (with affixes indicated by integers) with six unique markers illustrating the inverse correlation between i-complexity and e-complexity when number of markers is constant: (a) has relatively high e-complexity (1.58 bits) and low i-complexity (0 bits) , while (b) has relatively low e-complexity (0.83 bits) and relatively high i-complexity (0.52 bits). In (a) there are three different ways to mark each grammatical function (hence high e-complexity), and forms in all grammatical functions are predictive of all other forms (hence low i-complexity). In (b), on the other hand, there is only one realization for marking the plural number and two for marking dual (hence lower e-complexity), but in this organization the plural form is not predictive of forms in any other grammatical function and forms in dual do not fully predict the singular (hence higher i-complexity).

values (a maximum of 4.9 bits for Mazatec), while i-complexity stays relatively constant (between 0 and 1.1 bits). However, randomly generating paradigms of a fixed shape results in a similar distribution: e-complexity varies more than i-complexity[10], and when a paradigm has high e-complexity, it will necessarily also have low i-complexity. Ackerman and Malouf (2013) findings may therefore at least partly reflect the nature of the relationship between these two measures rather than anything specific to the dynamics of language change.

[10]Note however, that the paradigms generated here were matched in size to the paradigms used in Section 3.3 (3 inflectional classes and 3 grammatical functions); it could be that for much larger paradigms, such as found in natural languages, randomly generating the paradigms would result in higher i-complexity than values that can actually be found in natural languages (as suggested by the simulation with Chiquihuitlàn Mazatec done by Ackerman and Malouf (2013)).

3.4.3 The effects of i- and e-complexity on LSTM neural networks

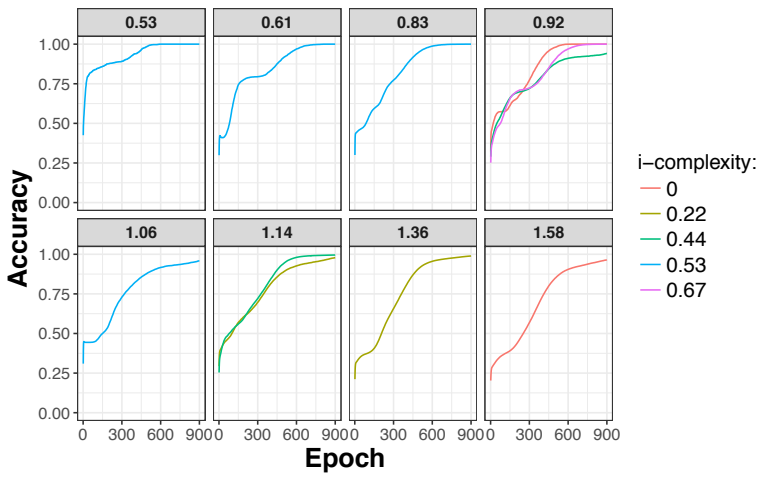
The learning results presented in section 3.3 already suggest that i-complexity has less impact on learning than e-complexity in networks, and possibly no impact in humans. To strengthen this conclusion, we also test how the 1000 randomly generated paradigms described above are learned using LSTM neural networks with the same architecture and parameters described in Section 3.3.2. Since the effects we found above held across networks of different sizes, here we only used networks of size 25 (4,656 parameters). We generated 50 different runs for each paradigm. In each run the initial weights of the network were randomly generated. As before, stems were randomly assigned into one of the three noun classes. Below we analyse accuracy in each epoch as well as the summed accuracy across epochs.

Results

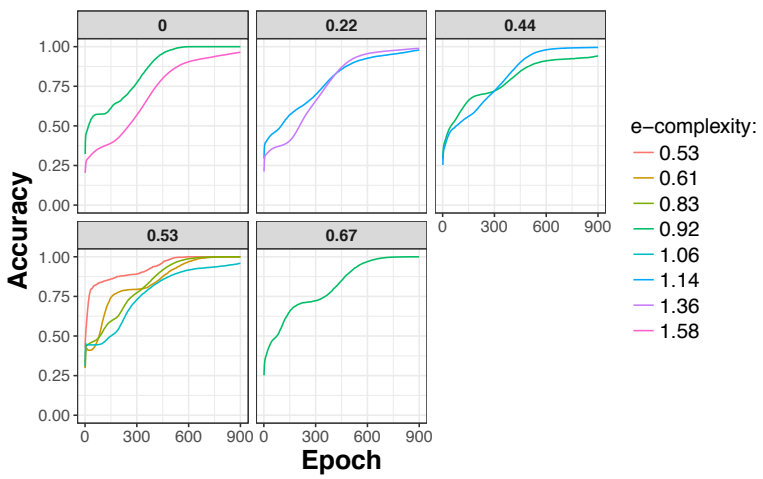
Figure 3.9 shows the learning trajectories of the neural networks in choosing the correct affix for lexemes, both by the i-complexity of the paradigm, and by its e-complexity.

To test how varying values of i-complexity and e-complexity affect learning, we ran a linear mixed-effects regression model predicting summed accuracy by paradigm i-complexity, paradigm e-complexity, the number of different affixes in the paradigm, and their interactions.

Summed accuracy was divided by 900 (number of epochs) to get the proportional summed accuracy, ranging from 0 to 1. I-complexity and e-complexity were scaled and number of markers was centred such that estimates for the effects of i-complexity or e-complexity reflect their effect on learning when the number of affixes equals the mean value (six affixes). In addition to these fixed effects, the model included random intercepts for different runs of the network (recall that network size was held constant).



(a)



(b)

Figure 3.9: Network learning trajectory for paradigms varying in i-complexity and e-complexity values. (a) i-complexity varying by color (facet titles showing e-complexity in bits). (b) e-complexity varying by color (facet titles showing i-complexity in bits). Note that, as discussed above, for some values of i-complexity, the random paradigms do not vary in e-complexity. In these cases, only one learning curve is shown (e.g., for e-complexity of 0.53 bits, there are only paradigms with i-complexity of 0.53 bits). Differences in e-complexity produce higher variability in network learning trajectories (b) compared to differences in i-complexity (a).

The model revealed a significant effect of both i-complexity ($\beta = -0.0093, t(49992) = -9.96, p < 0.001$) and e-complexity ($\beta = -0.04, t(49992) = -40.66, p < 0.001$). These

results replicate our initial findings with only four paradigms: increasing either the i-complexity or e-complexity of the paradigm leads to slower learning. Note that this holds even though, as discussed above, i-complexity and e- complexity have a strong inverse correlation ($r = -0.94$). Importantly, as before the effect size of e-complexity is much higher than the effect size of i-complexity (-0.04 vs. -0.009; approximately 4 times greater), suggesting a stronger effect of e-complexity on learning.

The model also reveals a significant effect of number of affixes ($\beta = 0.007, t(49992) = 18.51, p < 0.001$). Surprisingly, this effect is positive: more unique affixes appears to facilitate learning. However, a closer look at paradigms with the same i- and e-complexity and the same number of markers reveals a potential confounding factor, namely syncretism type. Table 3.6 shows an example of two of the random paradigms (labelled (a) and (b)), both of which have i-complexity of 0 bits, e-complexity of 1.58 bits, and 5 unique affixes (represented by numbers). While the proportional summed accuracy for paradigm (a) is 0.538, for paradigm (b) it is 0.87.

	Singular	Dual	Plural
noun class 1	1	2	8
noun class 2	8	3	5
noun class 3	3	8	1

(a)

	Singular	Dual	Plural
noun class 1	1	8	1
noun class 2	0	5	0
noun class 3	2	2	8

(b)

Table 3.6: Two example paradigms (with affixes indicated by integers) differing only in their degree of cross-class syncretism: (a) shows only across-class syncretism, while (b) shows mostly within-class syncretism. For both paradigms i-complexity (0 bits), e-complexity (1.58 bits) and number of markers (5 markers) are matched. Paradigm (b) is learned more accurately by our networks.

In paradigm (a), markers are distributed such that there is syncretism targeting forms across different noun classes. For example, the affix *1* marks singular for noun class 1, but plural for noun class 3. On the other hand, syncretic affixes in paradigm (b) are largely within noun classes. For example, the affix *1* marks singular and plural for noun class 1. There is one case of across-class syncretism in paradigm (b) – the affix *8* marks dual for noun class 1 but plural for noun class 3 – whereas in paradigm (a) there are 4 such cases. The learnability disadvantage for across-class syncretism is expected based on the previous results reported above. However, it turns out to lead to the unexpected apparent advantage for paradigms with more unique affixes, since paradigms with fewer affixes will tend to have more across-class syncretic forms in our design. We added number of across-class syncretic forms (centred) as a predictor in our previous regression model, including its interaction with the original predictors. This model again reveals a significant effect of i-complexity ($\beta = -0.0086, t(49992) = -9.12, p < 0.001$) and e-complexity ($\beta = -0.024, t(49992) = -23.42, p < 0.001$). The model also reveals a significant *negative* effect of number of affixes ($\beta = -0.034, t(49992) = -91.4, p < 0.001$), and a significant effect of the number of across-class syncretic forms ($\beta = -0.039, t(49992) = -151.1, p < 0.001$). Here, both of these effects are in the expected direction: having more unique affixes or having more across-class syncretic forms both lead to slower learning.

3.5 Discussion

In this study, we compared how different features of morphological paradigms affect learnability of morphological systems. Specifically, we compared measures reflecting the number of inflection classes in the paradigm and the number of different variants to mark each inflection (e-complexity), measures capturing the implicative structure of the paradigm and the

extent to which forms in the paradigm predict each other (i-complexity), number of affixes used in the paradigm, and type of syncretism (within versus across class). We tested the effects of these features on learning inflection paradigms with human participants and with recurrent neural networks (LSTMs). In Section 3.3 we compared the learnability of four artificially constructed nominal inflection paradigms differing either in e- or i-complexity. We found that changing the i-complexity of the paradigm had an effect on learning only in LSTMs but did not show an effect on learning in human participants. By contrast, e-complexity was found to have a stronger effect on learning in LSTMs relative to i-complexity and low e-complexity was beneficial for human learners. These results replicate the effects reported in Johnson, Culbertson, et al. (2020) and extend them to a more realistic learning scenario where input includes all forms at all staged (rather than restricting early input to predictive forms).

It is worth noting that the differences in i-complexity between our low- and high- complexity paradigms were not very large – the difference is 0.222 bits. It could be that larger differences in i-complexity values would reveal a larger effect on learning. However even this difference corresponds to complete predictability of the dual given the singular in the low complexity paradigm, compare to at best 66% predictability in the high complexity paradigm. In other words, while the difference as measured in bits is small, the difference in probability of correct prediction in the paradigm is large. Furthermore, the same size difference in e- complexity values did reveal a significant effect on learning. Testing more extreme values of i-complexity and e-complexity is in principle possible, but would necessitate training participants on much larger inflectional paradigms. This is challenging with human participants, since our experiment was already at the upper end of what we believe participants will tolerate in a single sitting; using the same methods for larger paradigms would probably necessitate a

multi-day experiment.[11]

Type of syncretism was also found to be predictive of learning in LSTMs; a paradigm with across-class syncretism in which the same affix is used to mark two different categories (e.g., singular and plural) for lexemes from separate inflection classes was learned slower than a paradigm with within-class syncretism, where the same affix is used to mark different numbers for lexemes within the same inflection class. This effect of syncretism on learning in LSTMs was seen both in Section 3.3, with the two example paradigms differing by types of syncretism, and in Section 3.4, when training the neural networks on paradigms with varying number of across-class syncretic forms. These results are compatible with studies with human learners showing that certain types of syncretism patterns are easier to learn than others (e.g., Maldonado and Culbertson 2019; Pertsova 2012). However, in our experiment with human learners, there was no effect of type of syncretism. Given the different results in the LSTMs and human learners, these mixed results call for a more systematic investigation into the effects of syncretism type on learning morphological paradigms.

Recall that Ackerman and Malouf (2015) suggested that morphological paradigms come to have restricted values of i-complexity through the process by which language users solve the Paradigm Cell Filling Problem for unknown forms. In other words, the mechanism by which i-complexity is kept low in natural language is generalization, rather than learning more generally. In Johnson, Culbertson, et al. (2020), we tested the effect of i-complexity on generalization with LSTMs, and our results there match Ackerman & Malouf's prediction: we saw a clear generalization advantage for low i-complexity paradigms. Together with our finding that i-complexity does not robustly affect paradigm learning in the absence of

[11]It is also worth noting that we only tested adult learners, and thus the scenario is most similar to adult L2 acquisition. It is of course possible that child L2 learners might behave differently, or that the effect of i-complexity is only relevant for first language acquisition. Although we have not specific reason to believe this is the case, one could in principle investigate child learners using the kind of study we have reported here.

generalization to completely novel forms, these results suggest that i-complexity may indeed influence how paradigms evolve, but primarily (or perhaps even solely) through its impact on generalisation.

However, this interpretation is made somewhat less plausible by the results from Section 3 investigating randomly generated paradigms. These results suggest that the low i-complexity that Ackerman and Malouf (2013) observed may to some extent reflect an intrinsic relationship between the two measures. Specifically, we found that for randomly-generated paradigms, e-complexity and i-complexity are strongly negatively correlated; crucially, there were no paradigms with both high e-complexity and high i-complexity (Figure 3.8). Moreover, the ranges of values the two measures exhibited were different, with lower and less varied values of i-complexity (0 to 1.667 bits) than the values of e-complexity (0.528 to 1.585 bits). Following these results from Section 3.4, we would therefore *expect* to find similar trends in natural languages, as indeed shown in Ackerman and Malouf (2013). Any typological observation deviating from this trend would call for a theoretical explanation.

In addition to manipulating e- and i- complexity, the number of affixes used in the random paradigms was not fixed and varied randomly from 3 to 8 affixes. This allowed us to test the effect of the number of affixes on morphological learning by the networks and to explore the relationship between this aspect of the paradigm and the two complexity measures. Number of affixes was found to positively correlate with e-complexity and to negatively correlate with i-complexity; an inflectional paradigm with low i-complexity is more likely to have a high number of affixes and to be more e-complex. Note that this gives support to our decision to use average cell entropy to measure e-complexity in this study; it is positively correlated with number of affixes in the paradigm, a common measure for e-complexity in the literature, in randomly generated paradigms.

The high inverse correlation between e-complexity and i-complexity was also found when

looking at a subset of paradigms with the same number of unique affixes (six). Together with the previous finding, showing that both e-complexity and i-complexity correlate with number of affixes, these results suggest that the inverse correlation between i-complexity and e-complexity derives from both the number of affixes in the paradigm, and from the way the affixes are organized in the paradigm; intuitively, when there is a frequent form with which a grammatical function is realized across noun classes, the entropy of this grammatical function is reduced and thus the overall e-complexity is likely to be lower. However, forms in this grammatical function are less likely to predict other forms in the paradigm and therefore its overall i-complexity is likely to be high. This is more likely to occur with low number of unique affixes in the paradigm, but the relationship between e- and i-complexity can be seen even when controlling for number of affixes.

Finally, generating the random paradigms also enabled us to test the effect of e- and i-complexity on learning with LSTM networks for larger range of values of the two measures, as opposed to the specific values we tested in Section 3.3. Again, we found that both e-complexity and number of affixes strongly predict learnability of the paradigm. I-complexity was also found to predict the learnability of the paradigm, but with a much smaller effect size (-0.0086 vs. -0.024 for e-complexity).

The strong effect of e-complexity (measured as average cell entropy) on the learnability of morphological paradigms found here suggests that the frequency of forms play an important role in the learnability of the paradigm. This is a further evidence for the pervasiveness of the effects of frequency on language learning (e.g., Ambridge et al. 2015). In the context of inflectional complexity, Sims and Parker (2016) suggest that in addition to implicative structure (i-complexity), type frequency of inflection classes also plays a role in reducing the complexity of the paradigm. In our experiments, type frequency of all noun classes was kept constant (with three words per noun class), but our results support the general claim that

the frequency of elements in the paradigm plays a role in inferring the correct inflected form for a lexeme.

To summarize, our findings suggest that a number of factors affect the learnability of inflection paradigms. However, these factors do not all play equal roles in determining ease of learning. The i-complexity of a paradigm does affect learning, at least in neural networks. But it is a relatively weak predictor of learnability relative to e-complexity (and number of unique affixes). Moreover, all paradigm features examined here were found to be interdependent, most crucially e- and i-complexity. This suggests that conclusions about the contribution of different types of complexity to natural language paradigms must take into account how measures of complexity relate to one another; observing measures independently can lead to potentially misleading conclusions about how different types of complexity might shape language.

Lastly, it is worth returning to the observation that e-complexity varies widely in morphological paradigms across languages. Since our findings show that e-complexity better predicts the learnability of the paradigm, all other things being equal, paradigms with low e- complexity should be preferred. Of course, learnability is not the only factor shaping linguistic systems: languages are used for communication, and linguistic systems have been claimed to reflect a trade-off between inductive biases (e.g., for simplicity) and pressure from communication (e.g., minimizing ambiguity, Kemp and Regier (2012)). This trade-off has been shown in a variety of linguistic domains, where natural languages show a near optimal balance between these two pressures (e.g., Regier et al. 2015; Xu et al. 2016; Zaslavsky et al. 2020). Evidence for this trade-off has also been found in experimental studies manipulating the relative importance of learning and communication (e.g. Kirby, Tamariz, et al. 2015; Motamedi et al. 2019; Silvey et al. 2015). Since we showed here that e-complexity correlates positively with number of distinct forms in the paradigm (i.e., distinctions in the lexicon),

morphological paradigms with high e-complexity could in principle reflect a balance between the communicative needs of speakers and the inductive biases of learners. Relatedly, it may be that e-complexity interacts with frequency effects coming from other aspects of the morphological paradigm and the lexicon. E-complexity captures the distribution of forms for each grammatical number, and thus reflects only the frequency of a specific aspect of the morphological paradigm. It is possible however that paradigms with high e-complexity have other means for reducing learning-relevant complexity, e.g. through skewed distribution of other aspects of the paradigm (e.g., inflection classes type/token frequencies or frequencies of forms of grammatical functions in the paradigm).

3.6 Conclusions

On the surface, natural languages exhibit a huge range of variation in terms of their inflectional paradigms; some languages have relatively little morphology, and others have large morphological paradigms with many inflectional classes, expressing many grammatical categories. How such large paradigms are acquired, and by extension how they persist across generations of learners is thus something of a mystery. A recent influential conjecture is that predictive structure is a shared feature of large paradigms one finds in natural languages (Ackerman and Malouf 2013). One possibility is that this predictive structure influences how languages change over time: inflectional paradigms have evolved under a pressure for low i- complexity (a measure of predictive structure in paradigms), rather than a pressure for low e- complexity (a measure of paradigm size). Here we presented results from a series of experiments with neural networks and human learners which muddy this picture. First, we find relatively small effects of i-complexity on learning, but robust effects of e-complexity. Further, we find that in randomly generated paradigms, e-complexity and i-complexity are

negatively correlated; roughly speaking, as paradigms get bigger, they will necessarily have more predictive structure. Although it may well be that learners use predictive structure when it's all they have to go on, our findings therefore suggest that pressure from learning should tend to favour low e-complexity rather than low i-complexity.

Part III

Investigating how i-complexity interacts with phonological and semantic cues for class membership

Preface to Chapters 4 and 5

In the studies presented in Chapters 2 and 3, I tested the effect of i-complexity on learning inflectional paradigms both in neural networks and human participants, and how i-complexity affects generalization of the paradigm to novel items in neural networks.

Results suggest some evidence for the effect of i-complexity on learning, although weaker than effects of e-complexity, a second measure of morphological complexity, in both learners. In neural networks, e-complexity and i-complexity were found to affect learning the forms in the paradigm, with a greater effect of e-complexity. In human learners, while the effect of e-complexity on learning was robust and was found in both of the behavioural experiments testing its effect (chapters 2 and 3), results show weak evidence for the effect of i-complexity on learning, in only one out of the four behavioural experiments.

Results also show that LSTM neural networks perform better in generalizing the paradigm to novel items when trained on languages with low i-complexity (chapter 2). Together with findings from Seyfarth et al. (2014), my findings support the hypothesis of Ackerman and Malouf (2013) and Ackerman and Malouf (2015) that low i-complexity facilitates solving the Paradigm Cell Filling Problem (Ackerman, James P. Blevins, et al. 2009), i.e., guessing the correct inflected form for a lexeme based on another known inflected form of the same lexeme.

In this part of the thesis, I extend my results regarding the role of i-complexity in learning and generalizing inflectional paradigms in two aspects. First, I test the effect of i-complexity on learning inflectional paradigms in languages in which a subset of the nouns are phonologically or semantically marked for their class membership. Second, I previously tested the effect of i-complexity on generalization only with neural networks. Here, I train both neural networks and human participants on languages with low and high i-complexity and test its effect on

their ability to generalize the paradigm to novel items.

Author Contributions

I conceived and designed the experiments and simulations and collected the data, conducted the analysis and wrote the paper; Kenny Smith, Jennifer Culbertson and Hugh Rabagliati provided advice on the design of the experiments and data analysis, and commented on the paper.

Chapter 4

Phonological Cues for Class Membership

4.1 Introduction

In the experiments described in Chapters 2 and 3, we used artificial languages where noun class membership was not determined by the phonology or semantics of nouns. This was done intentionally, to control for an alternative learning mechanism; in the case where class membership can be determined by the noun's semantics or phonology, its inflected form could be predicted based on these cues rather than based on knowledge of other forms in the paradigm (as captured by i-complexity).

There are studies that suggest that categories can be acquired and generalized based on distributional cues alone (i.e., in the absence of direct phonological or semantic cues). Mintz (2002) show that adult learners use distributional information in the form of frequent frames to form a category for the middle word in the frame. For instance, if the nonce words frame *sook-X-runk* repeats in the input, then all words appearing in the position of X would form one category. Crucially, these categories are learned in the absence of any phonological or semantic cue for category on the middle word (i.e., words in the position of X did not share phonological or semantic features except for their shared environment).[1] Reeder et al. (2013)

[1]This process is also used as an explanation for how children form syntactic categories–in particular,

show that adult learners can use co-occurrence statistics (distributional information) to form categories. Participants were familiarized with (Q)AXB(R) sentences where Q,A,X,B and R are all categories of words and Q and R words did not appear in all sentences. They found that participants were able to determine the grammaticality of novel sentences when being exposed to enough examples from the language, i.e., learners are able to form word categories based on distributional information alone. Furthermore, specifically related to the implicative structure of inflectional paradigms, Seyfarth et al. (2014) show generalization of an inflectional class system based on predictive relations between forms in the paradigm (discussed in detail in Chapter 3). Since their items were English nouns with nonce suffixes, Seyfarth et al. (2014) suggest that the memory load in their task was low, therefore enabling paradigm learning without redundancy in cues.

However, semantic or phonological cues for class membership are often present in natural languages; in many languages, semantic and phonological features of nouns play a role in determining how nouns are classified, as in gender systems, numeral classifiers systems and so on (Dixon 1986; Lakoff 1987; Fraser and Greville G Corbett 2000; Senft 2000; Aĭkhenval d 2000). For instance, in Zande, a language spoken in Zaire and Sudan, nouns are assigned to one of four gender classes according to their social gender and animacy (see Table 4.1) (Greville G. Corbett 1991). Semantic cues can be found in all systems of noun classification (e.g., social gender), but in most systems they are not sufficient to account for all nouns in the language. Phonological cues on the other hand, are not fully deterministic in most systems, and thus less reliable, but can account for larger portion of nouns (Greville G. Corbett 2013). In Hebrew for example, nouns ending with /a/ or /et/ tend to be feminine nouns, while nouns ending with other consonants tend to be masculine nouns.

More importantly, semantic and phonological cues for class membership have been shown

nouns and verbs (e.g., Mintz 2003; Monaghan et al. 2005).

Table 4.1: Noun assignment in Zande (from Greville G. Corbett 1991).

Criterion	**Gender**	**Example**	**Gloss**
male human	masculine	kumba	man
female human	feminine	dia	wife
other animate	animal	nya	beast
residue	neuter	bambu	house

to facilitate learning of subcategories. A number of studies show that learners are able to form and generalize word classes with implicative structure (e.g., when inflected forms in the systems are predictive of one another) only in the presence of phonological or semantic cues indicating the class of the word (e.g., Braine 1987; Brooks et al. 1993; Frigo and McDonald 1998; Kempe and Brooks 2001; L. A. Gerken et al. 2009; Ouyang et al. 2012).

Frigo and McDonald (1998) show that when learning an artificial language where some studied items are marked with phonological cues that indicate subclass membership, learners can generalize to novel unmarked items, based on distributional information. They trained participants on an artificial language containing two classes. Nouns in each class were used with two indicators (similar to definite and indefinite articles). The indicators for each class differed from one another (*jai* and *quo* were the indicators for items in class I, and *fow* and *mih* were the indicators for class II items). In the systematically marked version of the language, 60% of the items in each class were phonologically marked (i.e., class I nouns ended with *ash* and class II nouns ended with *gor*). The remaining 40% of nouns in each class did not not share phonological features (unmarked items). In another version of the language, the unsystematically marked version, there was no systematic marking of class; 60% of the items were phonologically marked but their phonological marker was as likely to appear in class I as in class II. The remaining 40% of the nouns were unmarked, as in the previous version. A third version of the language served as a control condition, where none

of the nouns were phonologically marked.

Participants were trained on items in the language, used together with their two indicators. Their task was to learn the items they were exposed to, with their appropriate indicators, and to generalize the correct indicators to novel items, phonologically marked and unmarked. For unmarked novel items, they were introduced with the novel word together with one indicator and had to predict the other indicator for that word. Frigo and McDonald (1998) ran three experiments using this design, manipulating marker salience, frequency and position. They found that systematic phonological cues improve learning the correct stem-indicator pairings and enable generalization to novel stems; participants in the systematic marking condition were able to choose the correct indicator for novel words at a rate better than chance, even for unmarked items (with no accessible phonological cue for class membership).

The effect of systematic phonological cues on generalization was found only in their experiments where the phonological markers on the stems were of high salience (e.g., full syllable) and close in position to the indicator (e.g. marked as a prefix when the indicator precedes the noun or marked as a suffix when the indicator follows the noun). This suggests that salience and position of the phonological marker on the stem are important for facilitating generalization of the class system to novel unmarked items.

Based on their results, Frigo and McDonald (1998) suggested a model according to which learners first link the markers (i.e., the phonological or semantic cues) to individual indicators and only then are able to link together the indicators used with a given class. In their experiment, for example, participants first learn that items ending with *gor* take the indicators *jai* and *quo* and items ending with *ash* take the indicators *fow* and *mih*. Only after establishing this knowledge they were able to link the two indicators of each class together (e.g., items that take the indicator *jai* also take the indicator *quo* in other cases), independently of the phonological marking on the stem. Put another way, they suggest

a bootstrapping process where the phonological cues facilitate forming subcategories which enable the use of distributional cues (e.g., paradigm's implicative structure); using the distributional cues in turn facilitates generalizing the system to novel items, even when these are not phonologically marked. While Frigo and McDonald (1998) main findings are based on generalization of the paradigm to novel items, the model they propose could also be applied to the task of learning the forms in the language.

The considerations discussed above raise the possibility that i-complexity may have more of an effect on learning of inflectional paradigms, if class membership is overtly marked by semantic or phonological cues. Our method was based on that of Frigo and McDonald (1998), who showed that learning the studied items was facilitated when a subset of items in the language were systematically phonologically marked for class (systematically marked condition). According to their model, the presence of phonological cues enables forming categories as a preliminary step, which later facilitates linking inflected forms within each category. This could therefore enhance the effect of i-complexity on learning the forms in the paradigm. However, in their study they did not manipulate the implicative structure of the paradigm (predictive relations from one indicator to another for each noun were available to participants in all conditions). In this study, we manipulate both the i-complexity of the paradigm and the presence of additional cues for class. Therefore, we are able to test how these two factors affect learning. Our hypothesis is that i-complexity interacts with systematic marking of class membership such that the effect of i-complexity on learning would be found in languages where phonological or semantic cues for class membership are present. We test this hypothesis on both neural networks and human participants.

In addition to the learning task used in the behavioural experiments in Chapters 2 and 3, here we add a generalization task to test participants' ability to generalize the class system to novel stems. If learners are able to form class systems and generalize them only in the

presence of referential or phonological cues for class membership, as suggested by Frigo and McDonald (1998) and supported by other studies (e.g., L. A. Gerken et al. 2009), we would expect an interaction effect between the paradigm's i-complexity and the presence of phonological or semantic cues for class membership. However, results from Seyfarth et al. (2014) suggest that human learners are also able to generalize noun class to novel items in the absence of such cues, based solely on the implicative structure of the paradigm (i.e., its low i-complexity). If this is the case, we would see an main effect of i-complexity, regardless of whether additional cues to class are present.

This part of the thesis proceeds as follows. In chapter 4 we train and test both neural networks and human participants on artificial languages manipulating both i-complexity and the presence of phonological cues. Note that in previous chapters, our results from experiments with neural networks show an effect of i-complexity on both learning and generalization in the absence of phonological cues for items' class membership. These results from the neural networks therefore do not conform to Frigo and McDonald (1998) model; networks were able to form classes based on distributional information alone. In testing neural networks (section 5.1) our objective is to explore whether the presence of phonological cues facilitates forming classes and whether this in turn makes the low i-complexity more useful for the network in learning and generalization.

In section 4.4 we test the interaction of i-complexity and phonological cues for class membership on learning and generalization in human participants. To preview, our results suggest that participants were not sensitive to the phonological cues for class membership that they were trained on. Therefore, in chapter 5, we test the same hypothesis using semantic cues for class membership on the assumption that they might be more salient. We first run a pilot study to test whether human participants are indeed sensitive to the semantic cues for class membership we use. After verifying that participants were able to pick up on the

semantic cues in this task, we then run the full experiment with human participants testing our hypothesis.

4.2 Target Paradigms

We constructed four paradigms, manipulating the i-complexity of the paradigm or the presence of phonological cues for class membership. We trained both neural networks and human participants on these paradigms to test the effect of i-complexity in languages with phonological cues to class membership.

The basic paradigms consisted of fifteen CVCV nouns[2] randomly paired with meanings for human participants (see Section 4.4 below). The small lexicon size allows the system to be learned with reasonable accuracy by human participants in a short experiment. An additional set of nouns (15 nouns for the neural networks and 24 nouns for the human participants) was used to test generalization of the paradigm (see a detailed description in Sections 4.3.2 and 4.4 below).

Table 4.2 presents how the CVCV nouns in the language were generated and their allocation to the three noun classes in the two phonological cues conditions. We follow Frigo and McDonald (1998) and phonologically mark 60% of the nouns in each noun class (three out of five nouns). In the systematic phonological cues conditions, marked nouns in each noun class all share the same repeated vowel in the CVCV noun pattern (following Culbertson, Jarvinen, et al. 2019). For example, in the language in Table 4.2(a) below, marked nouns in noun class 1 all share the pattern C_1iC_2i, marked nouns in noun class 2 share the pattern C_1uC_2u, and marked nouns in noun class 3 share the pattern C_1eC_2e, with the first and

[2]The lexicon in this experiment includes more stems than previous experiments (15 vs. 9 stems). We used phonological marking on a subset of stems in each noun class and therefore needed more than three items in each noun class.

second C representing two different consonants. Unmarked nouns include the two vowels that were not used to mark any of the noun classes in the marked items. The patterns C_1aC_2o or C_1oC_2a are used for unmarked nouns in the example language in Table 4.2(a). In the unsystematic phonological cues conditions, marked nouns included a repeated vowel, but there was no one vowel cuing class membership. For example, in Table 4.2(b) below, marked nouns of each noun class could equally be of the pattern C_1iC_2i, C_1uC_2u or C_1eC_2e. Unmarked nouns were generated the same as in the systematic phonological cues condition.

Table 4.2: Example languages with systematic (a) and unsystematic (b) phonological cues (C representing any consonant).

	noun class 1	noun class 2	noun class 3
marked items	C_1iC_2i C_1iC_2i C_1iC_2i	C_1uC_2u C_1uC_2u C_1uC_2u	C_1eC_2e C_1eC_2e C_1eC_2e
unmarked items	C_1aC_2o C_1oC_2a	C_1aC_2o C_1oC_2a	C_1aC_2o C_1oC_2a

(a) systematic phonological cues

	noun class 1	noun class 2	noun class 3
marked items	C_1iC_2i C_1eC_2e C_1uC_2u	C_1uC_2u C_1iC_2i C_1eC_2e	C_1eC_2e C_1uC_2u C_1iC_2i
unmarked items	C_1aC_2o C_1oC_2a	C_1aC_2o C_1oC_2a	C_1aC_2o C_1oC_2a

(b) unsystematic phonological cues

Paradigm structure was identical to the previous experiments. The nouns in each class were inflected for three numbers: singular, dual and plural. Inflectional markers were seven CVC monosyllabic suffixes (-fel, -fob, -fir, -fam, -fut, -fon, -fik), all starting with -f- to facilitate stem-affix segmentation. These inflectional markers were randomly allocated to cells in each paradigm (for each run of the network, or each human participant) such that both paradigms shared the same e-complexity value (1.14 bits) but differed in i-complexity. In the low i-complexity paradigm, the singular form of a word predicts the dual form, while

in the high i-complexity paradigm it does not. Table 4.3 shows two example paradigms. In the low i-complexity paradigm (A), if a stem takes the marker -fir in singular, then it takes -fut in dual; if a stem takes -fob in singular, then it takes -fam in the dual. In contrast, in the higher i-complexity paradigm (B), there is not such regularity: nouns with -fir in the singular take either -fam or -fut in the dual. The i-complexity value for the low i-complexity language is 0.222 bits vs. 0.444 bits for the high i-complexity language. Note that the distinct plural forms in each paradigm serve to distinguish the three classes of nouns. Without distinct plural forms, the low i-complexity paradigm would have fewer classes than the high i-complexity paradigm.

Table 4.3: Example paradigm for low i-complexity (a) and high i-complexity (b) languages.

	Singular	Dual	Plural
noun class 1	-fir	-fut	-fon
noun class 2	-fir	-fut	-fel
noun class 3	-fob	-fam	-fik

(a) low i-complexity paradigm

	Singular	Dual	Plural
noun class 1	-fir	-fut	-fon
noun class 2	-fir	-fam	-fel
noun class 3	-fob	-fut	-fik

(b) high i-complexity paradigm

These four language types were used in training and testing for both LSTM neural networks and human participants. Training the two types of learners on these languages was done using a staged learning design; learners were first trained on the singular forms of the nouns in the language, after which they were exposed to both singular and plural forms, and finally inflected dual forms were included. We used the staged learning design to increase the chances of finding an effect of i-complexity; in low i-complexity paradigms, the dual forms could be predicted from the singular. Therefore, the critical trials in our experiments are the dual items. We first test how well learners learn the dual forms in the language. We

additionally test how well learners generalize the dual suffixes to novel items, given exposure to the form in either the singular or plural. We also present results for learning the singular and plural forms in addition to learners' performance on the critical dual trials.

4.3 Experiment 1: neural networks

4.3.1 Network Structure

We trained and tested LSTM networks of the same structure as in Chapters 2 and 3. We trained the model with a range of embedding vectors dimensionalities for the input layer and LSTM hidden layer dimensionalities (from 2-dimensional embedding vectors and 2-unit layer (224 parameters) to 25 (5,100 parameters)). For each paradigm and set of hyperparameters, 50 runs were produced.

4.3.2 Procedure

We trained the model on input-output pairs for 15 stems in the full paradigm (stem-suffix pairs for the three grammatical numbers). We then tested their accuracy at learning these pairings and at producing the correct dual suffix for 15 additional stems, for which the network was trained on the singular form but never the dual. For these stems the network was required to generalise to the dual based on its representation of the full paradigm and its exposure to the singular forms for those 15 stems.

In each run of the network, 30 lexical items were divided into two equal sets: 15 learning stems and 15 stems used for the generalisation test (i.e. dual-novel items). The stems were assigned to noun classes so that each noun class included 5 learning stems and 5 dual-novel

stems. 60% of stems in each class (3 out of the 5 learning stems and 3 out of the dual-novel stems) were phonologically marked according to the phonological cues condition (see Section 4.2 above) and the remaining 40% were unmarked.

In each run, the model was trained and tested on input-output pairs in three blocks, summarized in Table 4.4 below. In block 1, the network was trained and tested on singulars for all stems; in block 2 the networks was trained and tested on singulars for all stems, plus plurals for the 15 learning stems; in block 3 the network was trained on singulars for all stems, plus plurals and duals for the 15 learning stems. Finally, it was tested on the entire paradigm for learning stems and on singular and dual for dual-novel stems. Each block consisted of 300 epochs, each consisting of a single pass through the specified training set (randomized), followed by a pass through the specified test set (randomized). During testing, the network was given an input and had to generate an inflection. Our results show performance in the testing phase.

Table 4.4: Training and testing regime in the neural networks.

Block	**Epochs**	**Training**	**Testing**
1	300	learning stems - singular dual-novel stems - singular (30 items)	learning stems - singular dual-novel stems - singular (30 items)
2	300	learning stems - singular, plural dual-novel stems - singular (45 items)	learning stems - singular, plural dual-novel stems - singular (45 items)
3	300	learning stems - singular, plural, dual dual-novel stems - singular (60 items)	learning stems - singular, plural, dual dual-novel stems - singular, dual (75 items)

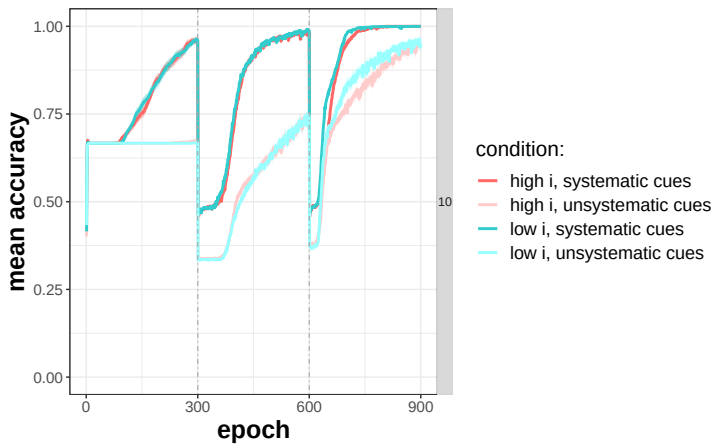
4.3.3 Results

Learning items

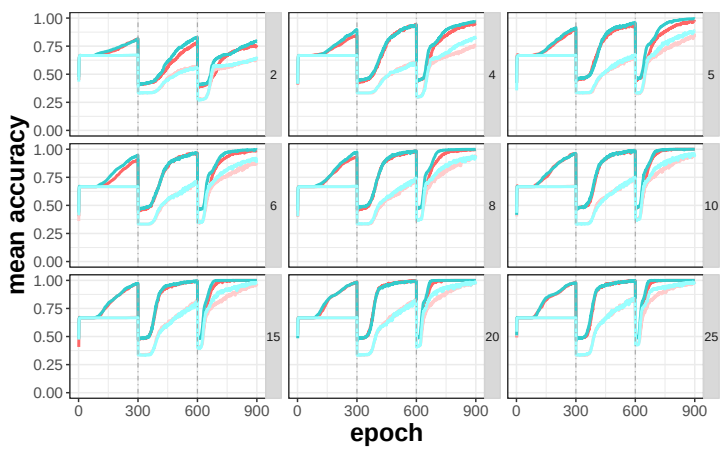
We measured the average accuracy of the networks in producing the correct affix for the learning items in all three grammatical numbers over epochs (averaged over 50 runs for each combination of target paradigm and network size). Fig. 4.1 presents the learning trajectories of the neural networks trained on the four languages over epochs. Networks of all sizes show higher accuracy levels for languages with systematic phonological cues (both low and high i-complexity) in learning the inflected learning forms. In the last 300 epochs (when networks are also trained and tested on forms in dual), networks trained on paradigms with low i-complexity show better performance compared to networks that were trained on high i-complexity paradigms and matching systematic cues condition. Networks of all sizes trained on languages with unsystematic cues for class membership, do not fully learn the forms in singular and plural prior to epoch 600 (where the dual forms are introduced), whereas almost all networks trained on languages with systematic cues (except for the two small sized ones) reach perfect or near perfect learning of the forms in singular and plural prior to epoch 600. By the end of training and testing (epoch 900), most of the networks, trained on both systematic and unsystematic cues languages, reach perfect or near perfect learning of all forms of the learning items (with the exception of very small sized networks - 2 to 6-dimensional embedding vectors and layer units).

To compare the difficulty of learning the languages (rather than whether the language is learnable or not), we compare the *mean summed accuracy* (i.e. the sum of the epoch-by-epoch accuracy rates divided by number of epochs) of the networks trained on the different languages. The mean summed accuracy reflects both the speed of learning and the accuracy throughout learning the language; in the results shown in Fig. 4.2, networks that learn the

language more rapidly have a higher mean summed accuracy.



(a)



(b)

Figure 4.1: Network learning trajectories for learning items inflected for singular, plural and dual. (a): results for one network size (10 cells), with line width indicating standard error, (b): results for all network sizes tested (facet titles give network size in number of cells). Vertical grey lines indicate the beginning of each block; at the beginning of Block 2 forms in plural are introduced to the networks and forms in dual are introduced in Block 3. Networks performance plunges at the beginning of each block of epochs as a result of the new forms they are introduced to. Networks trained on languages with systematic phonological cues (teal and red) show higher performance throughout the simulation. Networks trained on languages with low i-complexity and unsystematic phonological cues (light blue) show better performance than networks trained on the high i-complexity, unsystematic cues languages (pink) in Block 3, when duals are introduced.

we ran a linear regression model predicting the mean summed accuracy of the network in

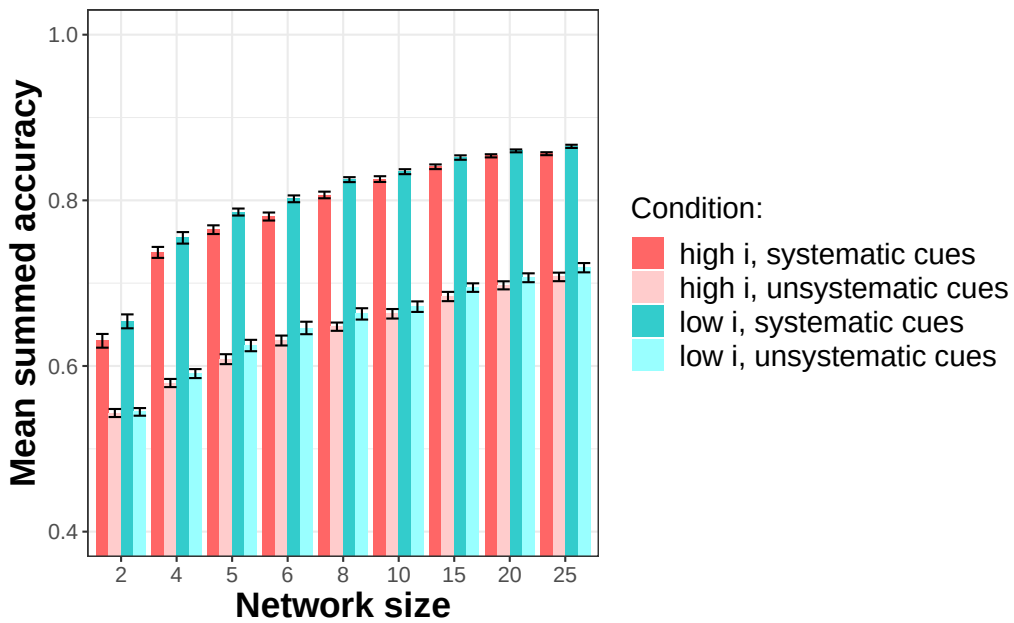


Figure 4.2: Mean summed accuracy on learning items over the 900 epochs of the networks trained on each of the four paradigm types across different sizes of the network.

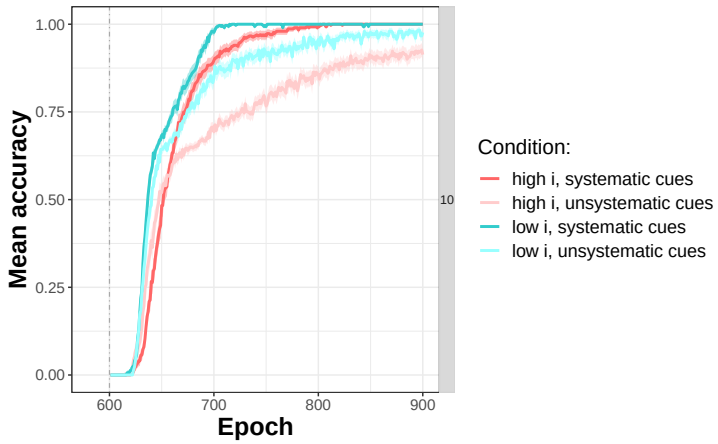
choosing the correct suffix for the learning items across all epochs, predicted by the language's i-complexity (high vs. low, sum coded), phonological cues (systematic vs. unsystematic, sum coded), size of the network (centered) and their interaction.[3] The model revealed a significant effect of phonological cues (b=0.07, t=64.5, p<0.001), showing that networks trained on languages with systematic phonological cues display faster learning with higher accuracy. The model also revealed a significant effect of i-complexity (b=0.006, t=5.6, p<0.001), confirming the better performance of networks trained on low i-complexity paradigms compared with their matched phonological cues condition, high i-complexity networks in the last 300 epochs, when the dual forms are introduced (see Fig. 4.1). Network size was also found to have an effect on learning accuracy (b=0.006, t=42.37, p<0.001), with large networks showing better performance throughout learning. The model also revealed a marginal inter-

[3] A linear mixed-effect regression model that included random intercepts for run number yielded a singular fit error. We therefore removed the random effects from the model and ran the linear regression model presented here.

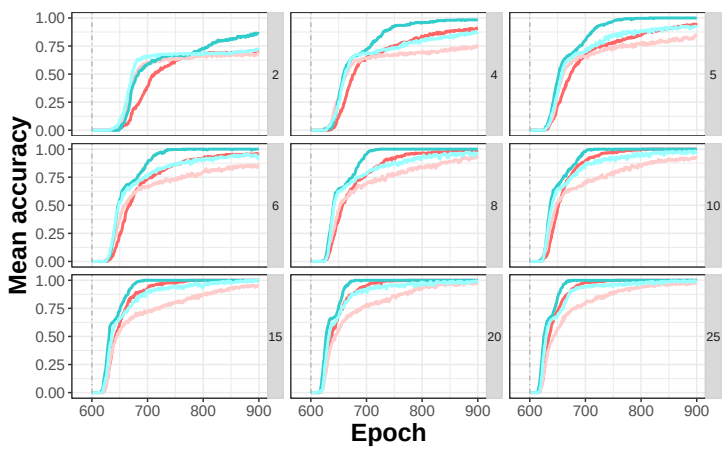
action between phonological cues and network size (b=0.0003, t=2.05, p=0.04), suggesting that larger networks benefited more from systematic phonological cues for class membership in learning. The model did not reveal an interaction between i-complexity and phonological cues (b=0.001, t=0.9, p=0.37). Note however that the critical items for testing the interaction between i-complexity and systematic phonological cues on learning are the studied dual forms, since the high and low i-complexity paradigms differ in the predictability of the dual form based on the form in singular. Within the dual forms, we are especially interested in learning the forms that are phonologically unmarked. Testing differences in learning these forms can reveal bootstrapping effects of i-complexity and systematic phonological cues, if it exists, since the unmarked dual forms lack the phonological marking to enable choosing the appropriate suffix for them directly based on this cue. In the low i-complexity paradigms their correct form can be predicted from their form in singular.

Fig. 4.3 presents the learning trajectories of the neural networks trained on the four languages in learning the studied unmarked dual forms over epochs (over epochs which include dual forms). Across all sizes of the network, except for networks with 2-dimensional embedding vectors and layer units, networks reach perfect learning of the dual forms by the end of the training-testing session (epoch 900). Across all network sizes performance was highest for networks trained on the low i-complexity, systematic language, and lowest for networks trained on the high i-complexity, unsystematic language. Fig. 4.4 shows the mean summed accuracy of all network sizes for the four languages in learning the forms in dual.

To test our hypothesis that systematic phonological cues interacts with i-complexity in learning the critical forms (unmarked studied items in dual), we ran a linear regression model predicting the mean summed accuracy of the network in learning the studied unmarked items in dual, from the language's i-complexity (high vs. low, sum coded), phonological cues (systematic vs. unsystematic, sum coded), size of the network (centered) and their interac-



(a)



(b)

Figure 4.3: Network learning trajectories in learning the studied unmarked items in dual (a): results for one network size (10 cells), with line width indicating standard error, (b): results for all network sizes tested (facet titles give network size in number of cells). Vertical grey lines indicate the beginning of Block 3, when the dual forms are introduced to the networks. Both low i-complexity and systematic phonological cues are shown to facilitate learning the unmarked dual forms, as the lowest performance is seen in networks trained on the high i, unsystematic cues languages (pink), and networks trained on low i, systematic phonological cues (teal) show highest performance across network sizes.

tion. This model showed the same effects of i-complexity, phonological cues and phonological cues and network size interaction (i-complexity: b=0.04, t=19, p<0.001; phonological cues: b=0.032, t=16, p<0.001; phonological cues and network size interaction: b=0.0008, t=2.99,

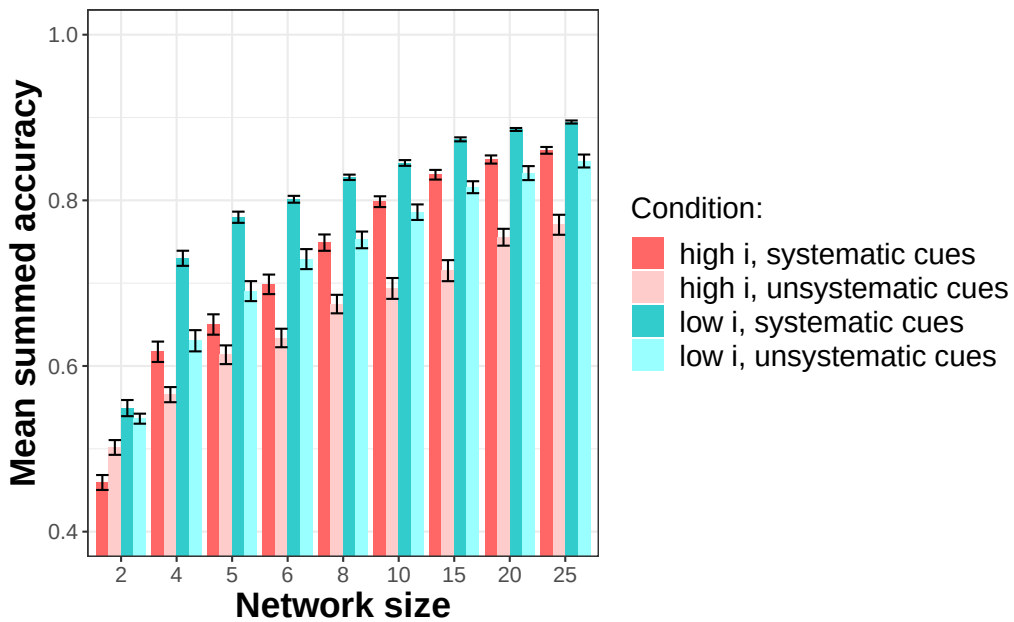


Figure 4.4: Mean summed accuracy in learning the unmarked items in dual over the last 300 epochs of testing, for networks trained on each of the four paradigm types across different sizes of the network.

p<0.01). In addition, i-complexity was found to negatively interact with size of the network (b=-0.0007, t=-2.7, p<0.01) suggesting that low i-complexity was more advantageous for small sized networks in learning the unmarked dual forms. Crucially, the interaction between i-complexity and phonological cues was non significant (-0.0006, t=-0.316, p=0.75). The model revealed however an interaction between i-complexity, phonological cues and size of the network (b=-0.0012, t=-4.41, p<0.001), suggesting that the interaction between phonological cues and i-complexity is different across sizes of the networks; in small networks, the effect of i-complexity on learning is greater in networks trained on languages with systematic cues, while in larger networks, the effect of i-complexity is greater in networks trained on languages with *unsystematic* cues. This interaction with network size was unexpected, and so we return to it in the Discussion.

These results suggest that the presence of systematic phonological cues for class membership

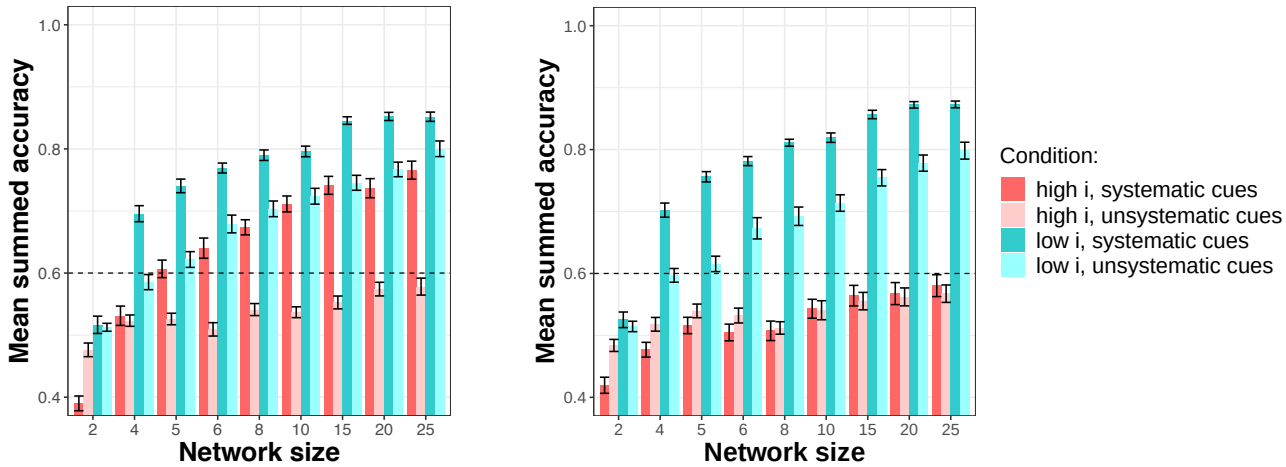
facilitates learning the forms in the language, even those that are not phonologically marked. However, the result do not provide evidence for the hypothesis that the presence of phonological cues further facilitates using the predictive relations between forms in low i-complexity paradigms for learning the dual forms; there is no clear evidence for interaction between i-complexity and phonological cues, as the interaction was found to differ across different network sizes. As in previous results from LSTM neural networks, however, i-complexity did affect learning.

Generalization

In the generalization task, we tested the networks' accuracy on dual forms with known singulars across the four language types. Fig. 4.5 shows the mean summed accuracy with which the networks chose the correct dual suffix for the novel items, separately for marked and unmarked items. Networks trained on languages with low i-complexity performed better than chance, both on marked and unmarked novel items in almost all sizes of the network. Performance was highest in networks trained on the low i, systematic cues language. For the large networks (>5), networks trained on high i-complexity, systematic cues languages performed better than chance on marked items, but lower than chance on unmarked items, as the unmarked stems do not include the phonological cue for class membership.

We ran a linear regression model predicting the mean summed accuracy of the network in generalizing the dual forms to novel items, from the language's i-complexity (high vs. low, sum coded), phonological cues (systematic vs. unsystematic, sum coded), size of the network (centered), item marking (marked vs. unmarked items, sum coded) and their interaction. As in previous models, the model confirmed a significant effect of i-complexity (b=0.08, t=51.7, p<0.001), and systematic phonological cues (0.034, t=20.82, p<0.001). The model also revealed an interaction between i-complexity and the size of the network (b=0.002, t=10.05,

p<0.001), suggesting that the effect of i-complexity is greater for larger networks (as opposed to the negative interaction found in learning the dual forms of unmarked studied items). There was also an interaction between phonological cues and size of the network (b=0.001, t=6.45, p<0.001), suggesting that larger networks benefited more from systematic phonological cues in generalizing the paradigm to novel items. The model also revealed a main effect of item marking (b=-0.013. t=-8.466, p<0.001), showing that accuracy was higher on marked items, an interaction between item marking and i-complexity (b=0.017, t=10.82, p<0.001), showing that effect of i-complexity on generalization was greater in unmarked items, and a negative interaction between phonological cues and item marking (b=-0.013, t=-8.47, p<0.001), showing that the effect of phonological cues was more pronounced for marked items. Crucially for our hypothesis, the model also revealed an interaction between i-complexity and phonological cues (b=0.01, t=6.27, p<0.001), confirming that in the case of generalization, i-complexity has a greater effect on generalizing the paradigm to novel stems when a subset of items in the language include phonological cues indicating class membership. The interaction between i-complexity, phonological cues and item marking was also found to be significant (b=0.017, t=10.74, p<0.0010), showing that the interaction between i-complexity and phonological cues differs across item marking; there is a bigger effect of i-complexity on generalization in languages with systematic phonological cues when looking at unmarked items (see Fig. 4.5). The model also revealed an interaction between i-complexity, phonological cues and size of the network (b=-0.001, t=-6.66, p<0.001), suggesting that the interaction found between i-complexity and phonological cues also differs across network sizes; the interaction between i-complexity and phonological cues is more evident in small networks, where having systematic cues in high i-complexity paradigms does not assist networks to achieve performance better than networks trained on the high i, unsystematic cues condition.



(a) Marked novel items in dual

(b) Unmarked novel items in dual

Figure 4.5: Mean summed accuracy in generalizing to novel marked (a) and unmarked (b) dual items in the last 300 epochs of testing, for networks trained on each of the four language types across different sizes of the network.

These results suggest that low i-complexity benefits generalizing the paradigm to novel stems more in languages where a subset of the items include an additional cue (phonological cue in this case) for class membership. These results are more apparent when looking at the specific items that are unmarked. This gives support to the hypothesis that phonological cues indicating class membership on a subset of items facilitate forming noun classes, which in turn makes the implicative structure in low i-complexity paradigms more advantageous in generalizing the paradigm to novel items.

4.3.4 Discussion

Altogether, results from the neural networks suggest that systematic phonological cues for class membership benefit learning the studied items. The effect of phonological cues on learning is seen also in *unmarked* items in dual. Systematic phonological cues were also advantageous in generalizing to novel forms in dual, and more so in marked novel items.

As in previous experiments (Chapter 2), low i-complexity was found to affect both learning and generalization in neural networks. The two factors, systematic phonological cues and i-complexity, were found to interact only in generalizing to novel forms, and not in learning; networks trained on languages with systematic phonological cues benefited more from low i-complexity when generalizing to novel forms.

Note that also in the case of generalizing to novel forms, where i-complexity interacts with phonological cues, low i-complexity benefits generalizing the paradigm independently of systematic phonological cues. These findings deviates from predictions drawn from Frigo and McDonald (1998)'s model.

Interestingly, while phonological cues were found to positively interact with size of the network, i-complexity was found to *negatively* interact with it. A three-way interaction (i-complexity * phonological cues * network size) was also significant and unexpected; in large networks the effect of i-complexity on learning was greater in languages with unsystematic cues. It may be that the three-way interaction results from the inverse two-way interactions of the size of the network with i-complexity and with systematic cues. The reason for i-complexity and systematic cues to interact differently with network size (i.e., to be more beneficial for learning and generalizing in different sizes of the network) is unclear. These findings call for further research. If similar patterns are apparent in other architecture and hyperparameters, it might reveal a principal difference between these two factors. Using the predictive structure of the paradigm (low i-complexity) or systematic cues on items in morphological learning are two independent strategies that are being used depending, in part, on the learner's processing power.

We next test the same hypothesis with human learners. As opposed to the neural networks, in previous experiments with human learners we found only weak evidence for the effect of i-complexity on learning. Here, we test whether i-complexity affects learning, when the

language includes phonological cues for class membership. In addition, we test whether i-complexity affects generalization of the paradigm to novel stems in dual, given the form in singular or plural.

4.4 Experiment 2: human learners

4.4.1 Methods

Materials

The same artificially constructed paradigms described in Section 4.2 were used to train and test human participants. Participants were exposed to the word forms in the language together with meanings. A set of 39 simple objects (animates: horse, frog, fox, cow, cat, monkey, hen, dog, shark, elephant, pigeon, giraffe; inanimates: ball, shirt, hammer, clock, glasses, comb, hat, bag, hand, shoe, crayon, bottle, guitar, bicycle, book, plane, broom, chair, spoon, lamp, mug, umbrella; botanical inanimates: lemon, orange, pear, tree, tomato) was randomly divided for each participant into two sets - a learning set of 15 items, and a set of 24 objects used for the generalization task.[4] Objects from the learning set were assigned to each of the three noun classes semi-randomly: each noun class consisted of two animate objects and three inanimate object, one of which is botanical. An example set of nouns assigned to the same noun class is [horse, frog, clock, ball, lemon]. This was done to ensure that noun class membership could not be determined based on semantic features. Stem-object pairing was done according to the phonological cues condition: for every participant,

[4]There are more generalization items in the experiment with human learners than with the neural networks (24 vs. 15) since here we were also interested in whether human learners are able to generalize the dual forms based on the form in plural. For the case of the neural networks, this was already seen in results from experiment in Chapter 2 (see Section 2.3.3).

3 objects in each noun class were randomly paired with marked stems, either providing a systematic cue to the noun class or not, and the remaining two were paired with unmarked items.

Objects from the generalization set were assigned to each of the three noun classes so that animate, inanimate and botanic objects were balanced across each of the three noun classes: an additional two animate and six inanimate objects, of which up to one is botanical, were assigned to each noun class. Half of the generalization items in each noun class (four objects) were randomly paired with phonologically marked stems, and half with unmarked stems.

Participants

203 self-reported native English speakers participants were recruited via Amazon's Mechanical Turk crowd-sourcing platform. They were compensated $8.5 for their participation and the experiment lasted 55 minutes on average (min = 24, max = 105, mode = 40). We recruited participants who possessed an Mturk qualification indicating that they were based in the US. Participants were allocated randomly to each of the four paradigms: low-i/systematic-cues (53); high-i/systematic-cues (48); low-i/unsystematic-cues (49); high-i/unsystematic-cues (53).

Procedure

The task consisted of two parts, learning the forms and generalization to novel stems. Learning the forms was done as in Chapter 2. Learning was staged over three blocks (singulars in block 1, singulars + plurals in block 2 and forms in all numbers in block 3). Since the lexicon in this experiment was larger than in Chapter 2, the current task included more trials (block 1: 60 trials; block 2: 120 trials; block 3: 180 trials). The order of trials was randomized in

each block.

A fourth block of trials formed the generalization phase of the task; in this block, participants were asked to choose the correct dual label for novel items. Dual trials of the novel items appeared after either a trial in singular or in plural of the same object. Participant were therefore in principle able to generalize dual forms to novel items either based on the phonological marking (for marked items in the systematically marked condition), or based on the inflected form of the stem in singular or plural. Fig. 4.6 represents two example trials in Block 4.

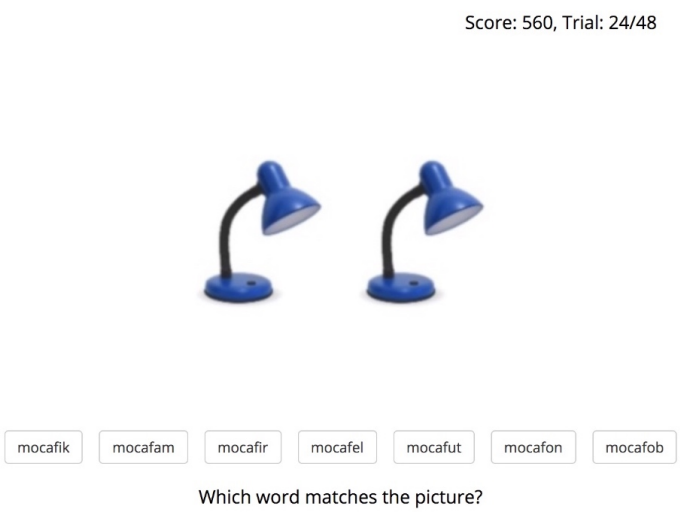
Block 4 consisted of 48 trials (two successive trials of each of the 24 novel items). Half of the dual trials of the novel items followed a singular trial and half a plural trial. Any effect of i-complexity on generalization would be predicted in dual trials which follow a singular trial. It is here that the paradigmatic relations between singular and dual forms differ across i-complexity conditions. However, we included trials in dual following a plural trial to test whether participants predicted the dual form from the plural (which is equally possible for both i-complexity conditions).

4.4.2 Results

Learning

Fig. 4.7 shows the mean accuracy with which participants chose the appropriate word form for singular and plural, as the experiment progressed trial by trial. On average, participants' accuracy was higher than chance in choosing the correct form in singular and plural throughout the task, suggesting that they learned the inflected forms in the language. Note that up to the end of Block 2, the task was identical for participants in the two i-complexity conditions (the difference between the high and low i-complexity conditions is introduced

(a)



(b)

Figure 4.6: Example pair of trials in the generalization block

with the dual forms in Block 3), varying only across phonological cue conditions. To verify that participants did not behave differently in the identical part of the task, we ran a mixed-effect logistic regression model predicting the accuracy in block two by i-complexity

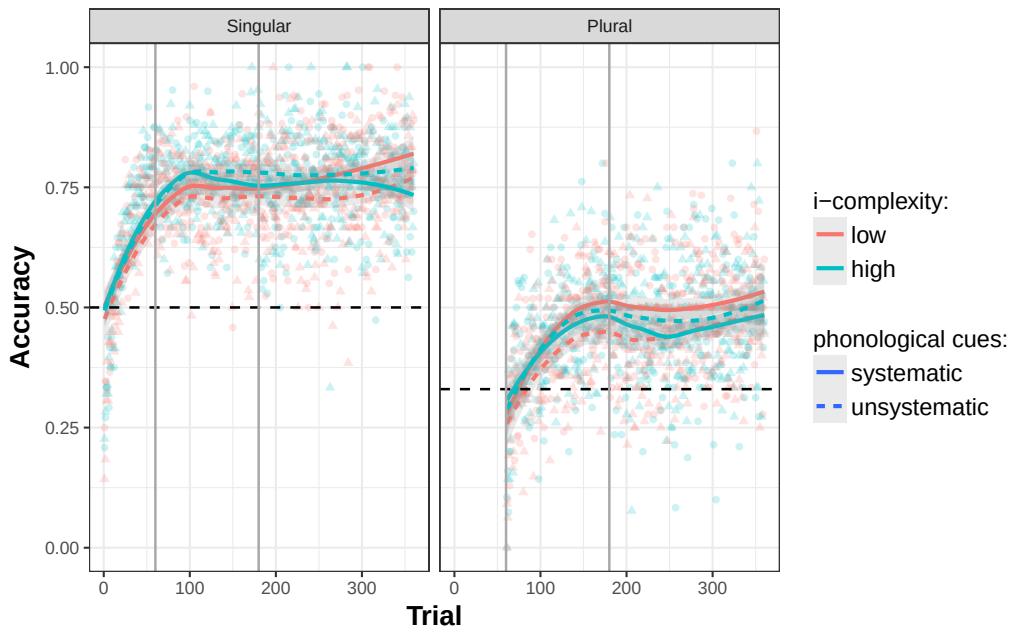


Figure 4.7: Mean accuracy by trial for singular and plural forms. Regression lines predicting accuracy by trial number for each of the conditions per grammatical number. Shaded points indicate mean accuracy scores averaged over participants in the systematic cues condition and shaded triangles indicate mean accuracy scores averaged over participants in the unsystematic cues condition. Horizontal dotted lines indicate the chance level for each form number (chance level is different for singular and plural forms according to the number of suffixes used to mark each number). Vertical grey lines indicate the beginning of each block; note that plural forms are introduced at the beginning of block 2. Participants in all conditions learned the singular and plural forms with accuracy higher than chance.

condition (high-i vs. low-i, sum coded), cue type (systematic vs. unsystematic, sum coded) and trial number (scaled).[5] The model also included by-participant intercepts and random slopes for trial number. The model revealed a significant effect of trial number (b=0.31, z=8.9, p<0.001), showing that participants' performance improved over time, but there was no significant effect of i-complexity (b=-0.06, z=-0.98, p=0.33) on performance in block 2, and no significant interaction between i-complexity and cue type (b=0.02, z=0.52, p=0.60) nor between i-complexity, cue type and trial number (b=0.04, z=1.2, p=0.23). This suggest that learners in all conditions were balanced with respect to their general ability to learn

[5]Model predictors were coded this way throughout unless otherwise noted.

in the task. Unexpectedly, there was also no significant main effect of cue type (b=0.037, z=0.62, p=0.53), suggesting that systematic phonological cues for class membership did not facilitate learning the forms in singular and plural.

We further test whether there is a difference between the marked and unmarked items in block 2. Fig. 4.8 shows mean accuracy of singular and plural trials in Block 2 for both marked and unmarked items. we ran a mixed-effect logistic regression model predicting accuracy in block two by item marking (marked vs. unmarked, sum coded), cue type, i-complexity condition and trial number. The model also included by-participant intercepts and random slopes for trial number. The model revealed a significant effect of trial number (b=0.31, z=8.57, p<0.001), but there was no significant effect of item marking on performance in block 2 (b=-0.02, z=-1.5, p=0.13) or an effect of cue type (b=0.035, z=0.6, p=0.55). There was also no significant interaction between item marking and cue type (b=-0.006, z=-0.44, p=0.66) nor an interaction between item marking, cue type and trial number (b=0.008, z=0.33, p=0.74). These results suggest that systematic phonological cues for class membership did not lead to a learning advantage, not even for the phonologically marked items. The lack of effect on learning can be due to participants not picking up on the phonological cues for class membership.

Fig. 4.9 shows the mean accuracy for dual trials in block 3, trial by trial. To test the effect of i-complexity and cue type and their interaction on learning, we ran a mixed-effect logistic regression model predicting accuracy in the dual trials in block 3 by complexity, cue type, trial number, participants' accuracy in block 2 (scaled) and item marking (marked vs. unmarked, sum coded). Item marking was included in the model to test whether any interaction between cue type and i-complexity holds across item marking or instead applies only to marked items. The model also included by-participant intercepts and random slopes for trial number. The model revealed a significant effect of trial number (b=0.69, z= 11.47, p<0.001), a significant

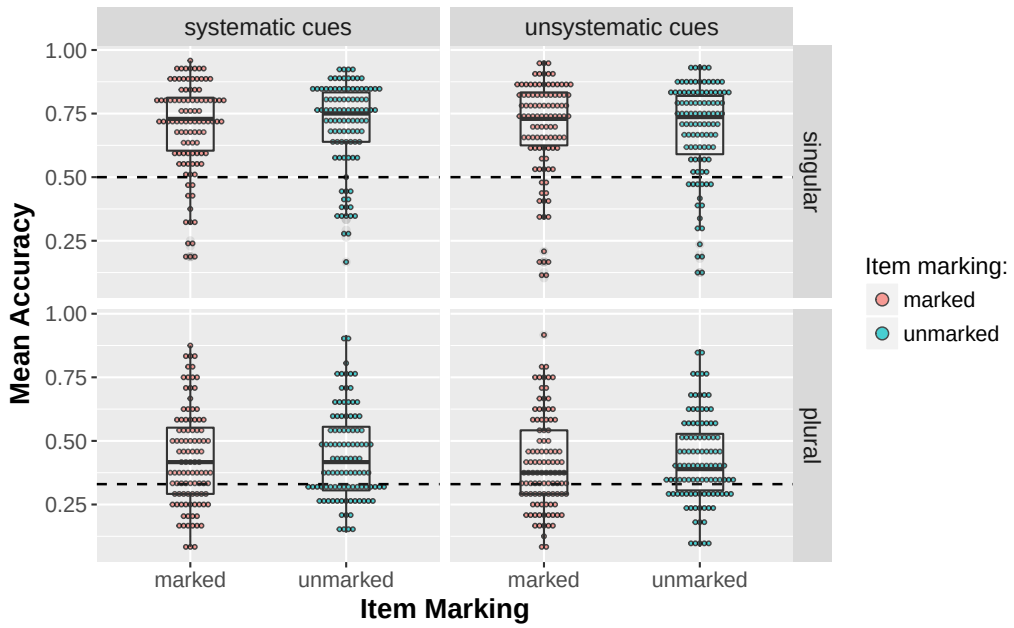


Figure 4.8: Mean accuracy for singular and plural trials by item marking. Points indicate each participant's mean accuracy scores in the systematic and unsystematic cues conditions (columns) separately for forms in singular and plural (rows). Horizontal line indicates chance level. Accuracy across conditions and across marked and unmarked forms is similar, suggesting that systematic cues does not facilitate learning the forms.

effect of accuracy in block 2 (b= 0.94, z=14.2, p<0.001) and a significant interaction between trial number and accuracy in block 2 (b=0.46, z=7.3, p<0.001) showing that participants' performance improved over time and participants who did well in block 2 were more likely to learn the dual forms in block 3 and to improve faster. The model also revealed a significant effect of i-complexity (b=0.34, z=5.2, p<0.001) as well as a significant interaction of i-complexity and trial number (b=0.12, z=2.1, p=0.036) showing that participants in the low i-complexity condition achieved higher performance in learning the forms in dual and their performance improved more over time than that of participants in the high i-complexity condition. Crucially, the model failed to reveal a significant effect of cue type (b=0.03, z=0.5, p=0.6) or an interaction between cue type and i-complexity (b=0.05, z=0.88, p=0.37) or between these two variables and item marking (b=0.02, z=0.89, p=0.37). These results

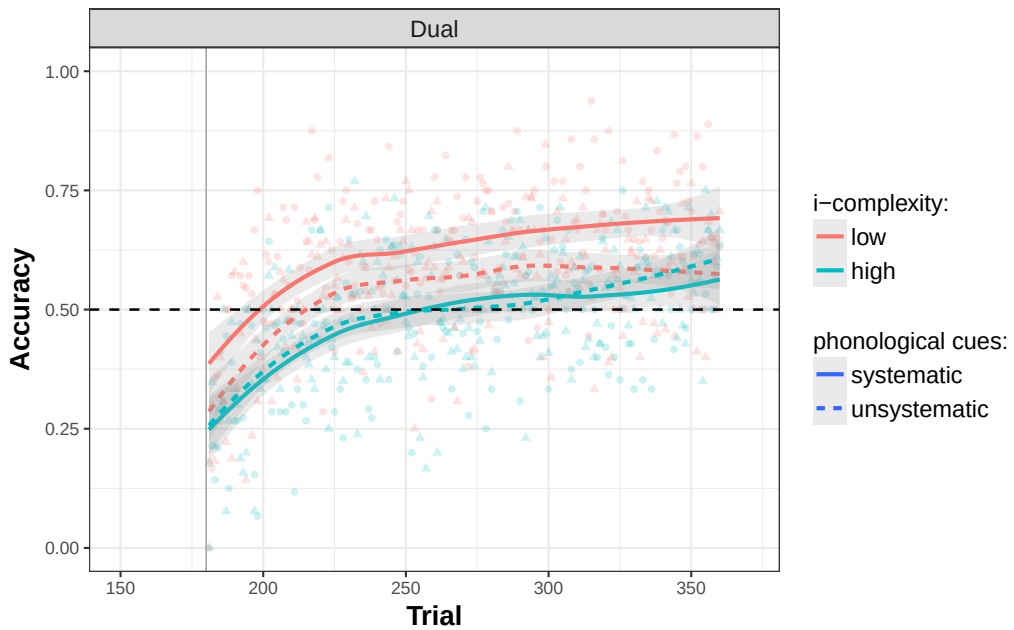


Figure 4.9: Mean accuracy by trial for dual forms. Shaded points indicate mean accuracy scores averaged over participants in the systematic cues condition and shaded triangles indicate mean accuracy scores averaged over participants in the unsystematic cues condition, with a regression line predicting accuracy by trial number for each grammatical number. Horizontal line indicates chance level. Vertical grey line indicates the beginning of block 3 when the dual forms were first introduced.

suggest again that learning the dual forms was not affected by systematical phonological cues to class present in the stems. Fig. 4.10 shows accuracy on dual trials by item marking. By comparison, in the simulation with LSTMs, the effect of phonological cues was already present in learning. This suggest that the lack of bootstrapping effect may be due to the fact that participants did not pick up on the phonological cues for class membership.

Generalizing to Novel Stems

Fig. 4.11 shows participants' generalisation accuracy for novel lexemes in singular and plural across cue type and item marking. Participants' accuracy is compared with chance level; the only way for participants to perform above chance in these trials is by using the phonological

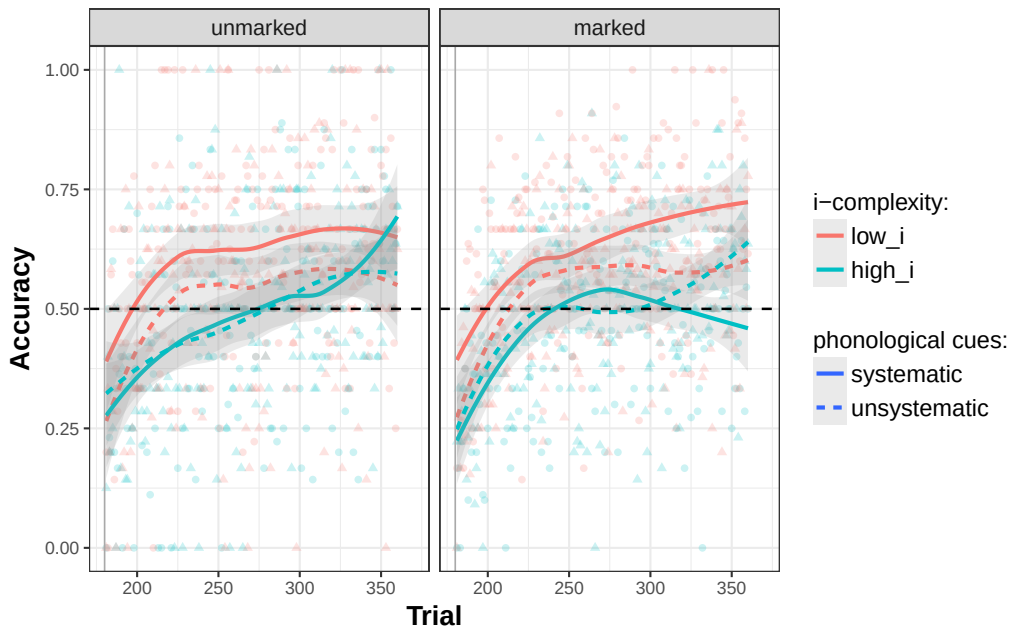


Figure 4.10: Mean accuracy by trial for dual marked and unmarked forms. Shaded points indicate mean accuracy scores averaged over participants in the systematic cues condition and shaded triangles indicate mean accuracy scores averaged over participants in the unsystematic cues condition, with a regression line predicting accuracy by trial number for each grammatical number. Horizontal line indicates chance level. Note that in conditions with no systematical phonological cues there is no difference between marked and unmarked forms.

cues on the stems. However, regardless of cue type, participants' performance is at chance, with no visible difference between marked and unmarked items. To test this statistically, we ran a mixed-effect logistic regression model predicting accuracy in singular and plural trials in the generalization phase (block 4) by complexity condition, cues type, item marking, participants' accuracy in block 2 and their interaction. For the novel items, we are not expecting learning over the small number of trial in block 4. Therefore we did not include trial number in this model.

The model revealed a significant effect of accuracy in block 2 (b=0.29, z=7.9, p<0.001) showing that participant who learned with higher accuracy the singular and plural forms in block 2 showed higher accuracy in generalizing the plural and singular marking to novel

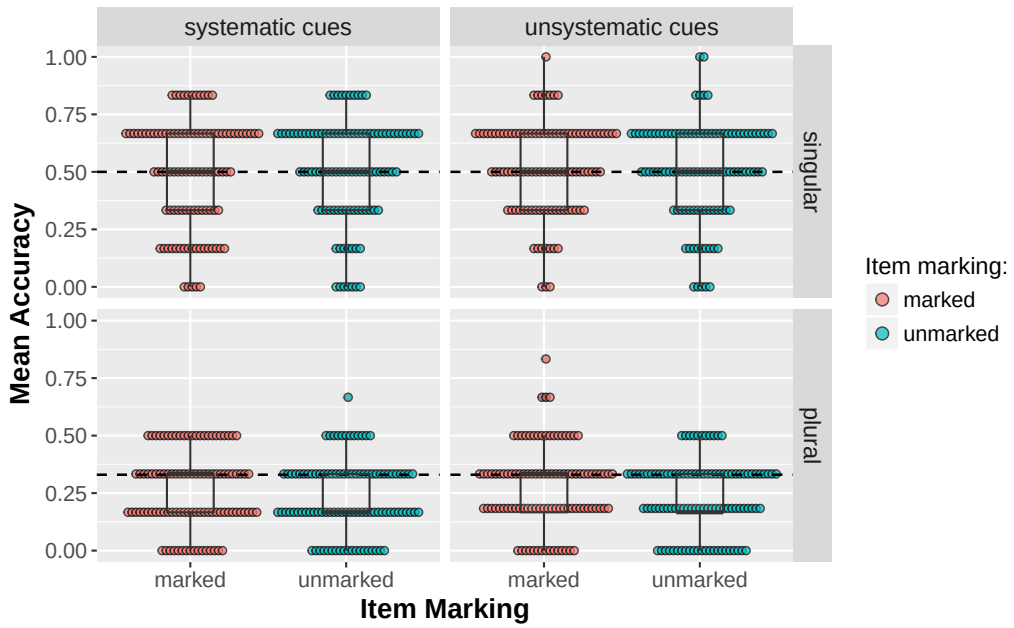


Figure 4.11: Mean accuracy for singular and plural trials by item marking. Points indicate each participant's mean accuracy scores in the systematic and unsystematic cues conditions (columns) separately for forms in singular and plural (rows). Horizontal lines indicate chance level. Note that chance levels reflect the number of suffixes marking each grammatical number rather than the number of different inflected forms participants can choose from in each trial; participants can perform lower than chance when choosing forms marked with suffixes for other numbers. The lack of difference in accuracy across item marking in the systematic phonological cues condition suggests that participants did not use the phonological cues for generalization.

stems. The model also revealed an interaction between accuracy in block 2 and cue type (b=0.084, z=2.28, p=0.023): better learners had an advantage in generalizing to novel stems in the systematic phonological cues conditions. This suggests that better learners did in fact pick up on the phonological cues for class membership. However, the model failed to reveal a main effect of cue type (b=-0.04, z=-1.08, p=0.28) or of item marking (b=-0.04, z=-1.35, p=0.17), confirming that among participants with average performance in block 2 (rather than only at the ones who learned better during block 2), systematic phonological cues did not aid generalization to novel stems for phonologically marked items. The model also failed to reveal an interaction between cue type and item marking (b=0.024, z=0.79,

p=0.43). Finally, as expected, there was no significant effect of i-complexity (b=0.06, z=1.61, p=0.1); at this point in the task, participants are not yet exposed to other forms of the novel stems. Thus forms cannot be predicted based on implicative relations in the low i-complexity paradigms yet.

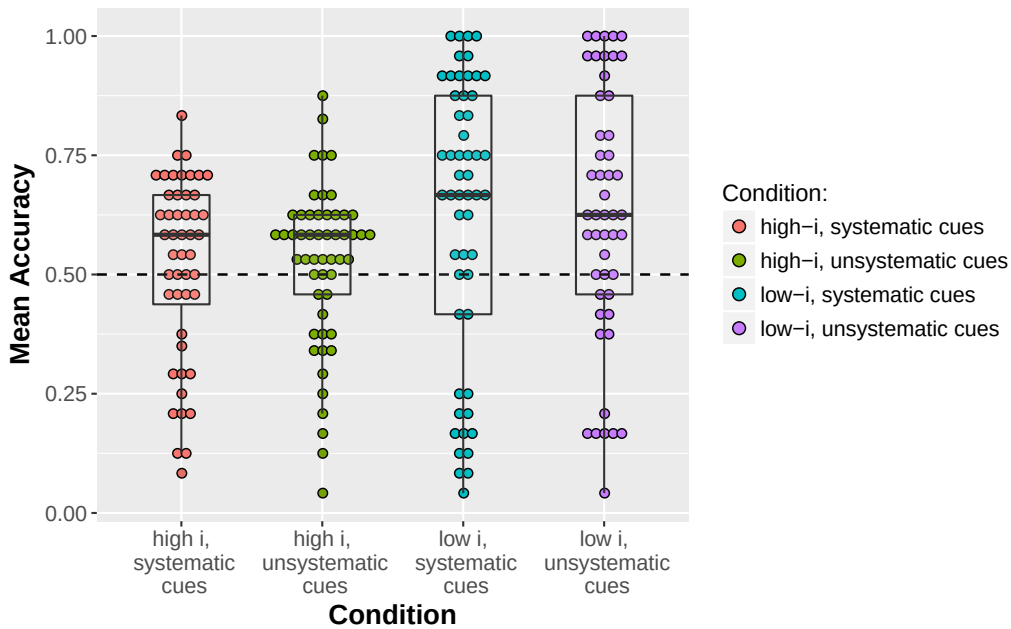


Figure 4.12: Mean accuracy for dual trials in the generalization phase for the four conditions (two i-complexity conditions and 2 phonological cues conditions). Points indicate each participant's mean accuracy with which they chose the appropriate form in dual. Horizontal line indicates chance level. Overall, participants in the two low i-complexity conditions show higher performance in generalizing new dual forms than participants in the two high i-complexity conditions. Cue type, however, does not affect participants' performance in generalizing to novel dual forms.

Fig. 4.12 shows each participant's accuracy in choosing the correct form in dual for novel lexemes, by condition. To test whether systematic phonological cues boosted the effect of i-complexity on generalizing the dual forms to novel stems, we ran a mixed-effect logistic regression model predicting accuracy in the dual trials in block 4 by complexity condition, cue type, participant's accuracy in block 2, item marking and the grammatical number of previous trial (singular vs. plural). Note that the grammatical number of the previous trial

was included as a fixed effect in the model since the difference in i-complexity between the high and low conditions differs only in the implicative relations between singular and dual forms. The model also included by-participant intercepts. Fig. 4.13 shows each participant's accuracy in generalizing the paradigm to novel forms in dual, split by item marking and the grammatical number of the previous trial. The model revealed a significant effect of i-complexity (b=0.35, z=5.04, p<0.001), as well as a significant effect of accuracy in block 2 (b=0.74, z=10.4, p<0.001) and a significant effect of the previous trial (b=0.1, z=3.1, p<0.01), suggesting that participants were better at generalizing the forms in dual after encountering the singular than after encountering the plural. The model also revealed a significant interaction between i-complexity and accuracy in block 2 (b=0.22, z=3.1, p<0.01) and a significant interaction between i-complexity and the previous trial (b=0.2, z=6.2, p<0.001), suggesting that the effect of i-complexity is greater in dual trials following trials in singular. The model failed to reveal an effect of cue type (b=-0.06, z=-0.8, p=0.4) or an interaction between cue type and i-complexity (b=-0.05, z=-0.7, p=0.48). These results suggest that contrary to our hypothesis and results from LSTMs, systematic phonological cues for class membership did not lead to an advantage in generalizing the paradigm to novel stems and did not enhance the effect of low i-complexity on generalization. The model also failed to reveal a significant effect of item marking (b=0.034, z=1.03, p=0.3) or an interaction between cue type and item marking (b=-0.007, z=-0.23, z=0.68), suggesting that systematic phonological cues did not improve generalization even for novel marked items. Taken together, these results are a strong indication that participants in this task have simply not noticed the phonological cues for class membership.

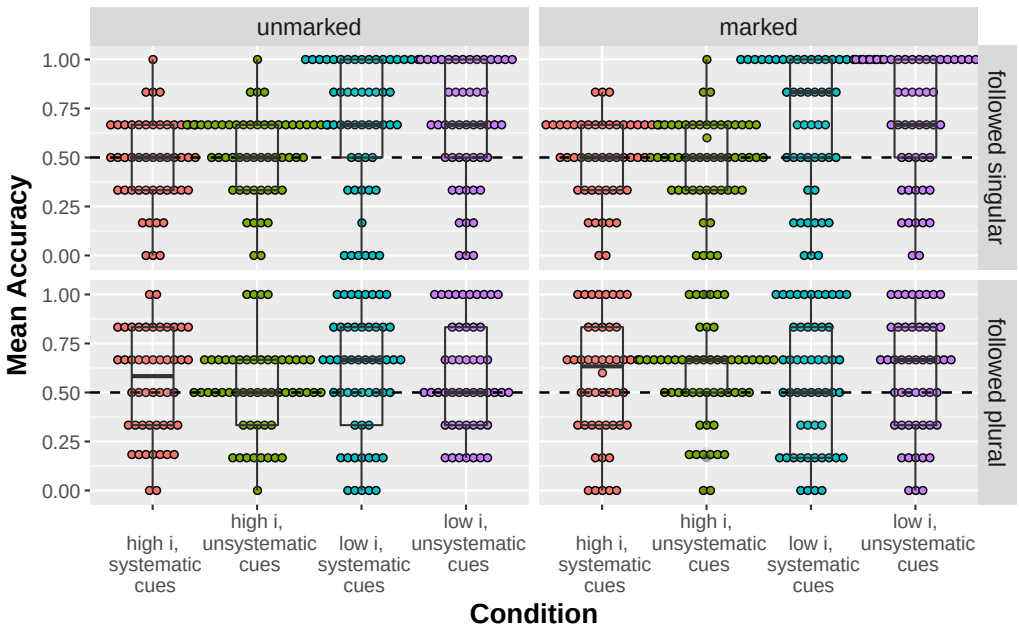


Figure 4.13: Mean accuracy for dual trials in the generalization phase split by item marking (columns) and the previous trial (rows). Dual trials that followed singular trials (upper row) show difference in performance across i-complexity conditions but not across cue type. This trend is seen both for the marked and unmarked items, suggesting that participants did not make use of the systematic phonological cues when generalizing to novel forms in dual.

4.4.3 Discussion

Results from learning and generalizing the forms in singular and plural suggest that participants did not pick up on the phonological cues in the language; there was no difference in participants' performance on marked vs. unmarked items in the systematic cues condition (Fig. 4.8, Fig. 4.11). Furthermore, in learning the dual forms and generalizing them to novel stems, there was no evidence for an effect of phonological cues, i.e., participants were not able to guess the correct dual form for a stem based on its phonological marker.

There are at least two possible reasons for the absence of an effect of phonological cues on learning and generalization in our task with human participants. First, in our task, words were presented orthographically rather than auditorily, with the expectation that learners

would read the words and notice the phonological patterns. However, this may not have been salient enough. Second, and not necessarily unrelated, Culbertson, Gagliardi, et al. (2017) show that adult learners favour semantic over phonological cues in a classification task where both are present. Therefore, adults may find such cues easier to learn. In Chapter 5 we therefore conducted another experiment using semantic cues for class membership.

However, it is also worth noting that participants in this experiment did apparently benefit from low i-complexity in learning and generalizing the dual. Our results therefore contribute to the body of evidence suggesting that inflectional classes can be learned independent of any extra-morphological cues for class membership, contrary to Frigo and McDonald (1998).

In previous studies, we saw weak evidence for the effect of i-complexity on learning. In only one out of three experiments with a staged learning design was there a significant effect of i-complexity. Here, the effect was found for both learning and generalization. Apart from the presence of phonological cues in this experiment, which our participants were evidently not sensitive to, the main difference in the design between this study and the experiments presented in Chapter 1, is the size of the language (15 vs 9 stems). To test whether the size of the language was driving the effect of i-complexity in learning, we conducted a combined analysis of data from the three experiments presented in Chapter 1, and data from the learning phase of the unsystematic phonological cues conditions from the experiment presented here. we ran a mixed-effect logistic regression model predicting accuracy in dual trials from i-complexity (high-i vs. low-i, sum coded), language size ('small' for the Chapter 1 experiments and 'large' for the experiment presented here, sum coded), participant's accuracy in block 2 (scaled) and trial number (scaled). The model also included by-participant intercepts and random slopes for trial number. The model revealed a significant effect of i-complexity ($b=0.29$, $z=5.15$, $p<0.001$), as well as an effect of trial number ($b=0.99$, $z=16.6$, $p<0.001$) and an effect of block 2 accuracy ($b=0.76$, $z=13.22$, $p<0.001$). The model also revealed a

significant effect of language size (b=-0.12, z=-21, p=0.036), suggesting that smaller languages are learned more easily. There was no significant interaction between i-complexity and language size (b=0.05, z=0.83, p=0.4). This suggest that there is no evidence that the larger size of the language is what led to the effect of i-complexity on learning. Rather, the effect of i-complexity may simply be present but weak.

Chapter 5

Semantic Cues for Class Membership

5.1 Experiment 1: pilot with human learners

In this chapter, we test the effect of semantic cues for class membership on learning and generalization. We first ran a pilot to test whether participants pick up on these cues, as opposed to the phonological cues we used in Chapter 4. The inflection paradigm used in the pilot experiment included only one grammatical number (singular). After verifying that participants can indeed use systematic semantic cues, Experiment 2 then tested the effect of i-complexity on learning and generalization in the presence of semantic cues for class membership (Section 5.2).

5.1.1 Methods

Materials

Participants were exposed to word forms in languages with semantic cues for class membership. As in Experiment 1, in Section 4.4, the language consisted of 39 CVCV nouns, however, in this experiment, nouns did not share phonological features and were randomly constructed. For each participants, nouns were randomly paired with 39 objects. Of the objects, 18 were inanimate (umbrella, broom, shoe, plane, clock, guitar, hat, bag, mug,

lamp, spoon, ball, chair, comb, shirt, bottle, glasses, book). These served as the semantically unmarked items. The remaining 21 were animates, split into three sub-categories: mammals (cow, giraffe, dog, cat, monkey, elephant, horse), birds (pigeon, parrot, seagull, owl, crane, eagle, swan) and insects (fly, rhino-beetle, grasshopper, ant, butterfly, red beetle, bee). These served as the semantically marked items. For each participant, the set of nouns was randomly divided into two sets, one including 15 nouns and comprising the learning set, and another of 24 nouns, comprising the set of novel items for the generalization task. Nouns from the learning set were allocated evenly into three noun classes. In each class, three out of five (60%) nouns referred to semantically marked objects from one of the three categories of animates (mammals, birds or insects). The remaining two nouns (40%) referred to randomly allocated unmarked inanimates. Table 5.1 shows an example allocation of nouns to the three noun classes.

Nouns in the language were inflected for singular according to their noun class. No other number inflection was used for this pilot. Inflectional markers were three CVC monosyllabic suffixes randomly chosen from the set [-fel, -fob, -fir, -fam, -fut, -fon, -fik] and randomly allocated, all starting with -f- to facilitate stem-affix segmentation.

Table 5.1: Example noun classification in a language with systematic semantic cues.

	noun class 1		noun class 2		noun class 3	
marked items	**birds**	pigeon seagull parrot	**mammals**	monkey elephant dog	**insects**	bee butterfly ant
unmarked items	hat shirt		bottle comb		lamp chair	

Objects from the generalization set were evenly assigned to each of the three noun classes (8 novel items in each noun class). Half of the novel items in each noun class were marked (animate) and were assigned to the appropriate noun class according to their semantics. The remaining novel items were unmarked items (inanimate) and were randomly assigned to the

noun classes.

Participants

21 self-reported native English speakers participants were recruited via Amazon's Mechanical Turk crowd-sourcing platform. They were compensated $3.5 for their participation and the experiment lasted 14 minutes on average (min = 11, max = 26, mode = 12). We recruited participants who possessed an Mturk qualification indicating that they were based in the US. Note that the manipulation (marked vs. unmarked items) was within participant, and all participants were exposed to systematic semantic cues.

Procedure

As in Chapter 4, the task consisted of two parts: learning, and generalization to novel stems. Recall that the language included forms in singular only, and so the generalization task tested whether participants had learned which types of animates went in which noun class. Learning was achieved via trial and error, as in previous experiments (see Chapter 2). The learning phase consisted of two identical blocks of trials (60 trials each), in which participants were exposed to all items in the learning set, 4 times each. The order of trials was randomized in each block and participants were allowed a self-paced break between blocks; they were presented with a screen announcing the end of the block and were asked to click on 'continue' to complete the next block of trials.

In the generalization task (block 3 with 24 trials), participants were asked to choose the correct label in singular for novel items (24 items to which participants were not exposed during learning, from the generalization set). To test whether participants picked up on the semantic cues for class membership, we looked at the accuracy for semantically marked

items compared with unmarked items, both in learning and in generalization to novel stems.

5.1.2 Results

Fig. 5.1 presents participant's accuracy on marked and unmarked items in block 2. To test whether participants were significantly better at learning marked items we ran a linear mixed-effect regression model predicting accuracy in each trial in block 2 by item marking (marked vs. unmarked, sum coded). The model also included a random intercept for each participant and a by-participant random slope for item marking. The model revealed a significant effect of item marking (b=0.44, z=3.4, p<0.001), showing that participants were better at learning items whose meaning indicated their class.

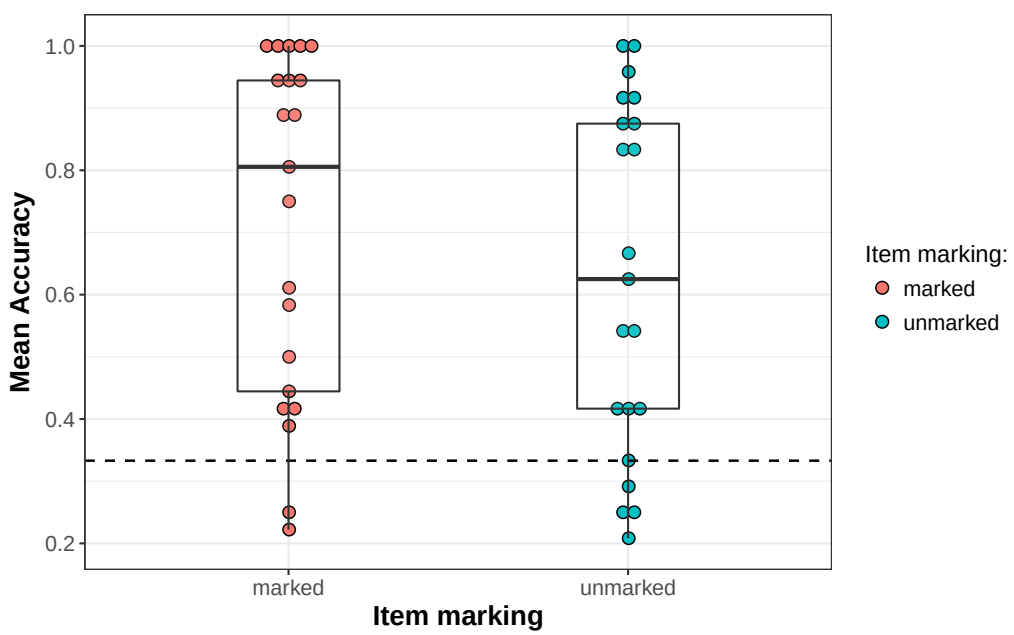


Figure 5.1: Mean accuracy for trials in block 2 by item marking. Points indicate each participant's mean accuracy over trials in the block. Horizontal lines indicate chance level. Participants' accuracy is higher on marked items (items whose meaning signals their class).

Fig. 5.2 shows participants' generalization accuracy for marked and unmarked items. We ran

a linear mixed-effect regression model predicting accuracy in generalization trials (block 3) by item marking, with random intercepts for each participant.[1] The model again revealed a significant effect of item marking (b=0.79, z=7.8, p<0.001), showing that participants were better at generalizing novel items whose meaning indicated their class. For the unmarked items, performance was at chance.

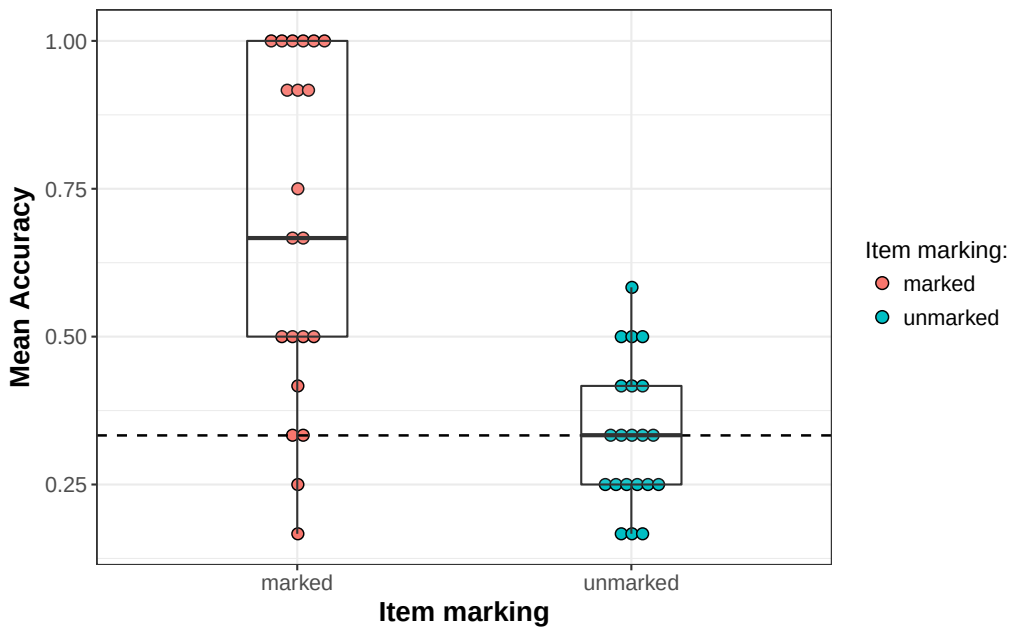


Figure 5.2: Mean generalization accuracy on novel items (block 3) by item marking. Points indicate each participant's mean accuracy over trials in the block. Horizontal lines indicate chance level. Participants' accuracy is higher for marked items.

Altogether, results from this pilot study show that semantic cues for class membership were advantageous for participants both in learning the forms and in generalizing the inflections to novel items. This confirms that participants pick up on the semantic cues for class membership and make use of them in learning and generalization. We now move on to test the effect of i-complexity on learning and generalization in the presence of this extra-morphological cue in a full experiment.

[1] A model that included a by-participant random slope for item marking produced a singular fit. We therefore ran the model without this random effect

5.2 Experiment 2: full experiment

5.2.1 Methods

Materials

As in Experiment 4.4, we constructed four conditions, crossing the i-complexity of the paradigm (high vs. low) and the presence of semantic cues for class membership in the language. Systematic semantic cues for class membership were constructed as in the pilot experiment (see Section 5.1.1). In this experiment we add a condition with unsystematic semantic cues, where semantically marked items are shuffled between noun classes so that noun class is not determined by animate object category. The learning set consisted of 15 nouns allocated to each of the three noun classes, see Table 5.2 for example noun allocation in the two semantic cues conditions.

An additional set of 24 nouns was used to test generalization of the inflection paradigm to novel items. Nouns from this set were evenly assigned to each of the three noun classes (8 novel items in each noun class). Half of the novel items in each noun class were marked (animate) and were assigned to the appropriate noun class according to the semantic cues condition (either assigned to the noun class matching their semantic category, in the systematic semantic cues condition, or randomly assigned to a noun class, in the unsystematic semantic cues condition). The remaining novel items were unmarked (inanimate) and were randomly assigned to the noun classes.

To test the effect of i-complexity on learning and generalization in the presence of semantic cues for class membership, we also manipulate the paradigms' i-complexity, as in Chapter 4. The nouns in each class were inflected for three numbers: singular, dual and plural. Inflectional markers were seven CVC monosyllabic suffixes (-fel, -fob, -fir, -fam, -fut, -fon, -

	noun class 1		noun class 2		noun class 3	
marked items	birds	pigeon seagull parrot	mammals	monkey elephant dog	insects	bee butterfly ant
unmarked items	hat shirt		bottle comb		lamp chair	

(a) systematic semantic cues for class membership

	noun class 1	noun class 2	noun class 3
marked items	pigeon	monkey	bee
	butterfly	seagull	elephant
	dog	ant	parrot
unmarked items	hat shirt	bottle comb	lamp chair

(b) unsystematic semantic cues for class membership

Table 5.2: Example languages for the systematic semantic cues condition (a) and the unsystematic semantic cues condition (b).

fik), all starting with -f- to facilitate stem-affix segmentation. These inflectional markers were randomly allocated to cells in each paradigm for each participant such that both paradigms shared the same e-complexity value (1.14 bits) but differed in i-complexity. In the low i-complexity paradigm, the singular form of a word predicts the dual form, while in the high i-complexity paradigm it does not. Tables 5.3 shows two example paradigms.

Table 5.3: Example paradigm for low i-complexity (a) and high i-complexity (b) languages.

	Singular	Dual	Plural
noun class 1	-fir	-fut	-fon
noun class 2	-fir	-fut	-fel
noun class 3	-fob	-fam	-fik

(a) low i-complexity paradigm

	Singular	Dual	Plural
noun class 1	-fir	-fut	-fon
noun class 2	-fir	-fam	-fel
noun class 3	-fob	-fut	-fik

(b) high i-complexity paradigm

Participants

205 self-reported native English speakers participants were recruited via Amazon's Mechanical Turk crowd-sourcing platform. They were compensated $8.5 for their participation and the experiment lasted 58.5 minutes on average (min = 21, max = 138, mode = 52). We recruited participants who possessed an Mturk qualification indicating that they were based in the US. Participants were allocated randomly to each of the four paradigms: low-i/systematical-cues (51); high-i/systematic-cues (53); low-i/unsystematic-cues (51); high-i/unsystematic-cues (50).

Procedure

The procedure was identical to the experiment with phonological cues, described in Section 4.4.1. We measured the accuracy with which participants chose the correct label for objects, in learning and generalization. Critical trials were the dual trials in both phases of the task. We also present participants' accuracy in singular and plural trials, to verify that participants picked up on the semantic cues for class membership and to test whether learning of these forms was higher than chance.

5.2.2 Results

Learning

Fig. 5.3 shows the mean accuracy on singular and plural over trials. On average, participants' accuracy was higher than chance throughout the task, suggesting that they learned the inflected forms in the language. Note that up to the end of Block 2, the task was identical for participants in the two i-complexity conditions (the difference between the high and

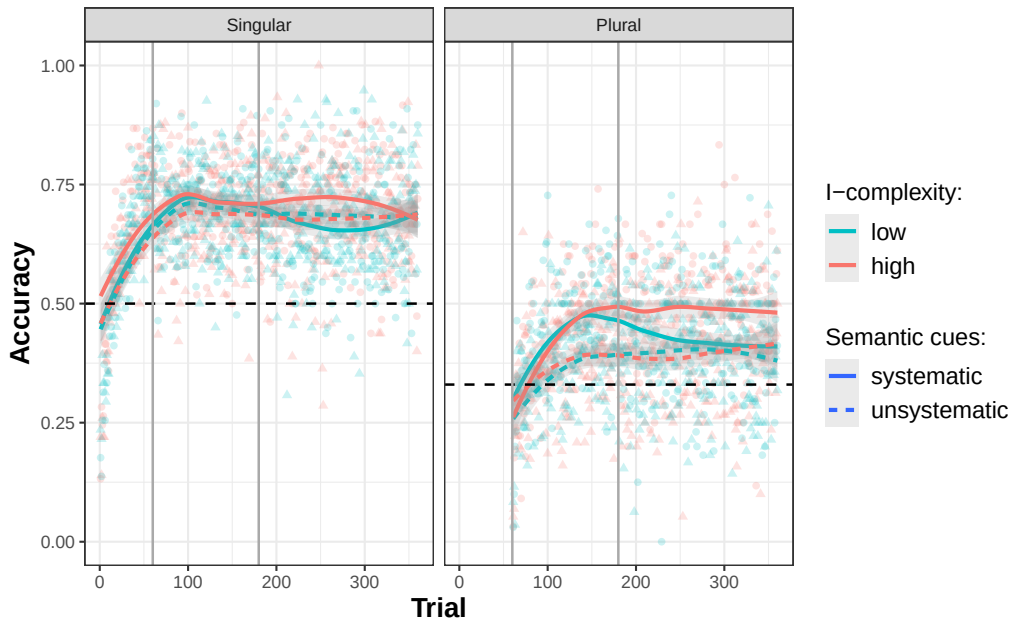


Figure 5.3: Mean accuracy by trial for singular and plural forms. Shaded points indicate mean accuracy scores averaged over participants in the systematic cues conditions and shaded triangles indicate mean accuracy scores averaged over participants in the unsystematic cues conditions, with learning trajectories averaged over participants. Horizontal lines indicate chance level for each number (chance level is different for singular and plural forms according to the number of suffixes used to mark each number). Vertical grey lines indicate the beginning of each block; note that plural forms are introduced at the beginning of block 2. Participants in all conditions learned the singular and plural forms with accuracy higher than chance. Participants across i-complexity conditions behave similarly in block 1 and 2, suggesting that participants were matched in terms of their general learning abilities across conditions.

low i-complexity conditions is introduced with the dual forms in Block 3), varying only across semantic cues conditions. To verify that participants did not behave differently in the part where the task was identical, we ran a mixed-effect logistic regression model predicting accuracy in block 2 by i-complexity condition (high-i vs. low-i, sum coded), cue systematicity (systematic vs. unsystematic, sum coded) and trial number (scaled).[2] The model also included by-participant intercepts and random slopes for trial number. [3] The model revealed

[2]Model predictors were coded this way throughout unless otherwise noted.
[3]All following models include these random effects unless noted otherwise.

a significant effect of trial number (b=0.27, z=7.3, p<0.001), showing that participants' performance improved over time, and a significant interaction between trial number and cue systematicity (b=0.098, z=2.62, p<0.01), showing that performance of participants in the systematic semantic cues condition improved faster. This effect is as expected, and suggests that participants picked up on the semantic cues for class membership. Furthermore, there was no significant effect of i-complexity (b=0.003, z=0.055, p=0.95) on performance in block 2, and no significant interaction between i-complexity and cue systematicity (b=-0.005, z=-0.08, p=0.93). The model therefore does not reveal a difference in performance across i-complexity conditions in block 2, suggesting that learners in both conditions were balanced with respect to their general ability to learn in the task.

We further test whether the there is a difference in learning the marked and unmarked items in the systematic cues conditions in block 2. Fig. 5.4 shows the mean accuracy in Block 2 by item marking. We ran a mixed-effect logistic regression model predicting accuracy in block 2 by item marking (marked vs. unmarked, sum coded), cue systematicity, i-complexity condition and trial number. As usual, there was a significant effect of trial number (b=0.27, z=7.04, p<0.001). Crucially, the model revealed a significant effect of item marking (b=0.065, z=4.56, p<0.001), and a significant interaction between item marking and cue systematicity (b=0.085, z=5.99, p<0.001). This shows that marked items were learned with higher accuracy, and the advantage of marked items was greater for participants in the systematic semantic cues condition. There was also a significant interaction between cue systematicity and trial number (b=0.08, z=2.1, p=0.036) and a significant interaction between item marking, cue systematicity and trial number (b=0.084, z=3.357, p<0.001), showing that performance of participants in the systematic semantic cues condition improved faster, and more so in trials with semantically marked items. As in the previous model, there was no significant effect of i-complexity on accuracy in block 2 (b=0.0026, z=0.042, p=0.966).

These results suggest that, as opposed the experiment with phonological cues, systematic semantic cues for class membership lead to a learning advantage in learning the forms in the language, especially the semantically marked items.

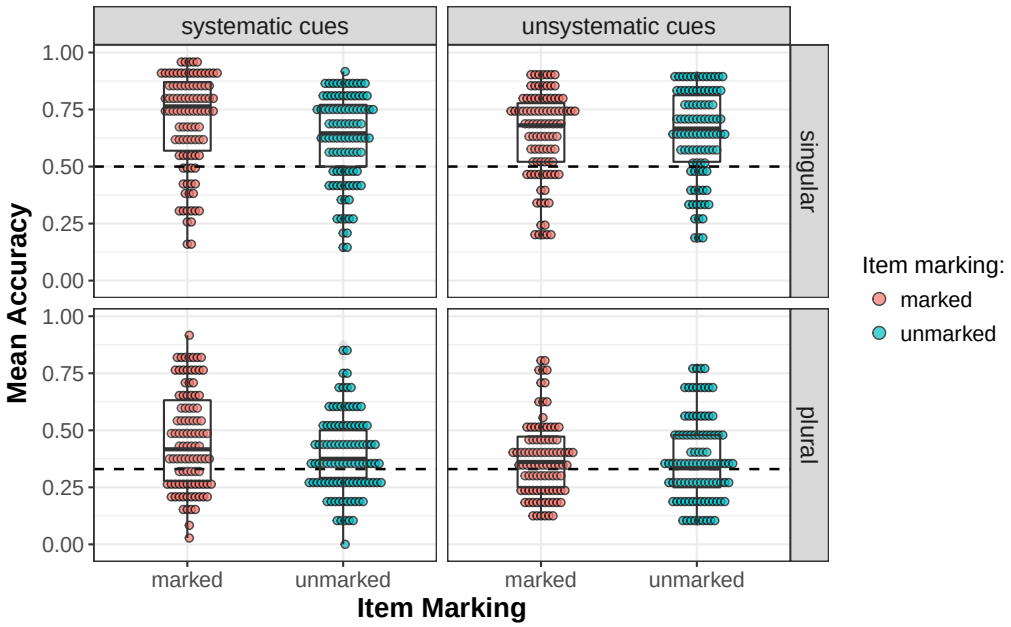


Figure 5.4: Participant's mean accuracy for singular and plural trials by item marking. Points indicate each participant's mean accuracy scores in the systematic and unsystematic marking conditions (columns) separately for forms in singular and plural (rows). Horizontal line indicates chance level. Accuracy in marked items (red) is higher in the systematic semantic cues conditions, suggesting that semantic cues for class membership facilitate learning the marked forms.

Fig. 5.5 shows accuracy on dual trials in block 3 by trial and item marking. To test the effect of i-complexity and semantic cues and their interaction on learning the dual forms, we ran a mixed-effect logistic regression model predicting accuracy on dual trials in block 3 by complexity condition, cue systematicity, trial number, participant's accuracy in block 2 (scaled) and item marking.

The model revealed a significant effect of trial number (b=0.54, z= 8.7, p<0.001), a significant effect of accuracy in block 2 (b=0.94, z=14.1, p<0.001) and a significant interaction

between trial number and accuracy in block 2 (b=0.302, z=4.53, p<0.001) showing that participants' performance improved over time and participants who did well in block 2 were more likely to learn better the dual forms in block 3 and to improve faster. The model revealed a marginally significant negative effect of cue systematicity (b=-0.12, z=-1.9, p=0.057). This unexpected negative effect suggests that participants in the systematic semantic cues conditions were worse at learning the dual forms. However, the model revealed a significant positive interaction between cue systematicity and item marking (b=0.046, z=2.12, p=0.034), and a significant interaction between cue systematicity, item marking and accuracy in block 2 (b=0.064, z=2.66, p<0.01). This suggest that participants in the systematic semantic cues conditions were better at learning the dual forms of semantically marked objects, especially for participants who showed higher performance in block 2.

The model also revealed a marginally significant effect of i-complexity (b=0.12, z=1.91, p=0.055) and a significant interaction between i-complexity and accuracy in block 2 (b=0.14, z=2.15, p=0.03), suggesting that participants in the low i-complexity conditions who showed higher accuracy in block 2, had higher accuracy in learning the dual forms.

Crucially for our hypothesis however, the model did not reveal a significant interaction between i-complexity and cue systematicity (b=-0.01, z=-0.15, p=0.88) or an interaction between i-complexity, cue systematicity and item marking (b=-0.002, z=-0.77, p=0.44). These results suggest that systematic semantic cues are generally advantageous for learning the dual forms of marked items, and similarly there is a general effect of i-complexity at least for better learners. However, the results do not provide evidence that these two factors interact; learning a language with semantic cues for class membership does not bootstrap use of predictive features in low i-complexity paradigms. Note that this same pattern was found for trained items in the RNNs with phonological cues (Section 5.1).

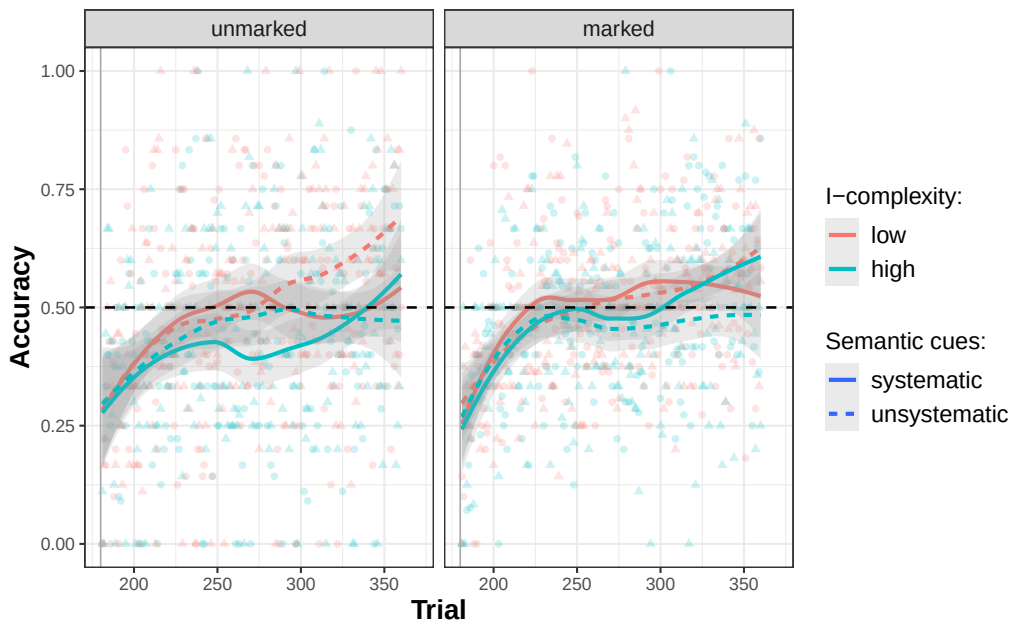


Figure 5.5: Mean accuracy by trial for dual marked and unmarked forms. Shaded points indicate mean accuracy scores averaged over participants in the systematic cues conditions and shaded triangles indicate mean accuracy scores averaged over participants in the unsystematic cues conditions, with learning trajectories averaged over participants. Horizontal line indicates chance level. Splitting the dual trials into marked and unmarked items reveals a moderate improvement in accuracy for participants trained on languages with systematic semantic cues in the marked items.

Generalizing to Novel Stems

Fig. 5.6 shows accuracy in choosing the correct form for novel lexemes in singular and in plural by cue systematicity and item marking. Participants' accuracy is compared with chance level as the only way for participants to choose the correct form for the novel lexemes in singular and plural is by using the semantic cues. On average, participants in the systematic semantic cues conditions are above chance at choosing the correct form in singular for novel semantically marked items. Along with the evidence above (Fig. 5.4) this suggests that participants picked up on the semantic cues for class membership. To test this statistically, we ran a mixed-effect logistic regression model predicting accuracy on singular and plural

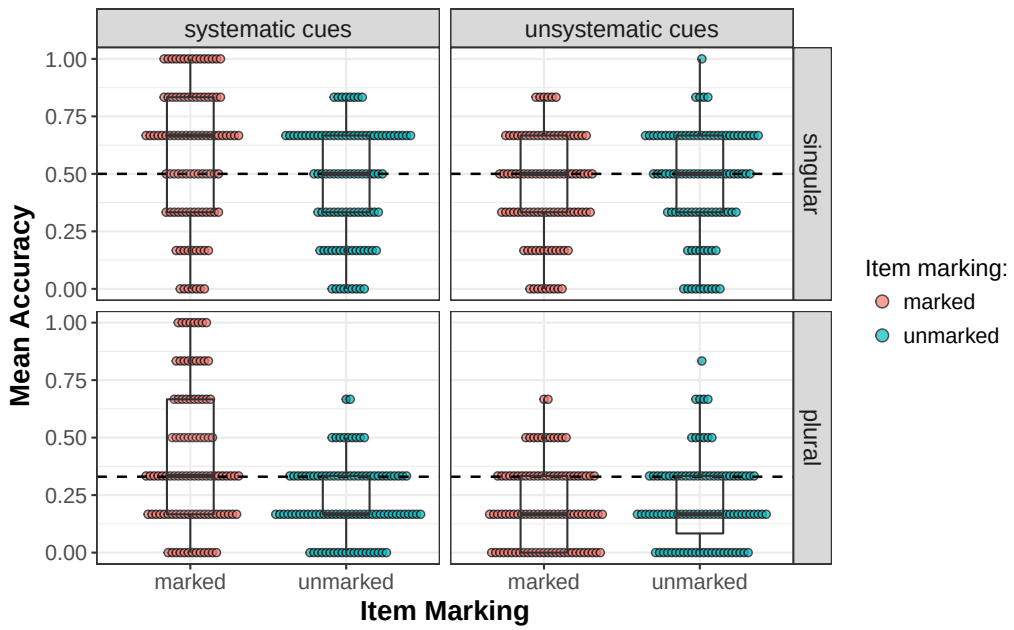


Figure 5.6: Mean accuracy for singular and plural trials by item marking. Points indicate each participant's mean accuracy scores in the systematic and unsystematic cues conditions (columns) separately for forms in singular and plural (rows). Horizontal lines indicate chance level. In the systematic cues conditions (left), accuracy in marked items (red) is higher than unmarked items (teal), both for singular and plural trials, suggesting that systematic semantic cues for class membership have a facilitative effect in generalizing to novel stems.

trials in the generalization phase (block 4) by complexity condition, cue systematicity, item marking and participant's accuracy in block 2 and their interaction. We did not include trial number in this model; given the short number of trials in block 4 we do not expect to see learning.

The model revealed a significant effect of cue systematicity (b=0.13, z=2.98, p<0.01), a significant effect of item marking (b=0.12, z=3.58, p<0.001) and a significant interaction between cue systematicity and item marking (b=0.17, z=5.16, p<0.001). This shows that participants trained on systematic semantic cues were more likely to generalize to novel stems in the singular and plural, and more so for semantically marked items. The model also revealed a significant effect of accuracy in block 2 (b=0.4, z=8.87, p<0.001), an inter-

action between accuracy in block 2 and cue systematicity (b=0.13, z=3.02, p<0.01), and an interaction between accuracy in block 2, cue systematicity and item marking (b=0.15, z=4.55, p<0.001). This suggests that the facilitating effect of systematic cues on generalization to novel stems was more accessible to participants who showed better learning in the learning phase. There was no significant effect of i-complexity (b=-0.06, z=-1.32, p=0.18).

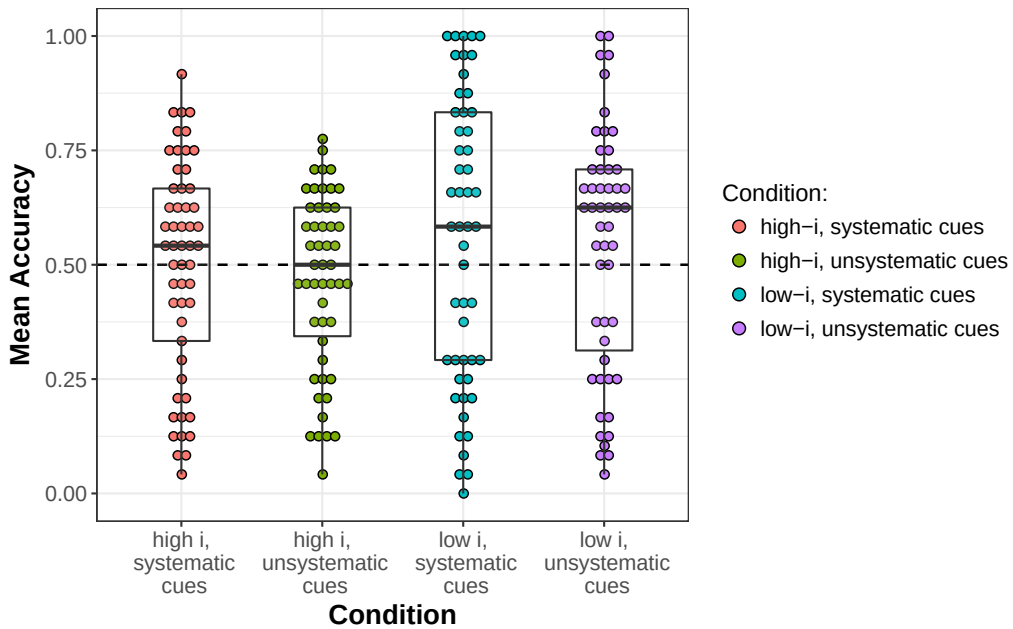


Figure 5.7: Mean accuracy for dual trials in the generalization phase for the four conditions (two i-complexity conditions and two semantic cues conditions). Points indicate each participant's mean accuracy with which they chose the appropriate form in dual. Horizontal line indicates chance level. Participants in the two low i-complexity conditions show higher accuracy at generalizing the dual forms to novel stems, although this difference is small. There is no clear difference between the two semantic cues conditions.

Fig. 5.7 shows accuracy on dual trials for novel lexemes by condition. To test whether systematic semantic cues boosted the effect of i-complexity on generalizing the dual forms to novel stems, we ran a mixed-effect logistic regression model predicting accuracy in the dual trials in block 4 by complexity condition, cue systematicity, participant's accuracy in block 2, item marking and the grammatical number of previous trial (singular vs. plural).

Note that the grammatical number of the previous trial was included as a fixed effect in the model since the difference in i-complexity between the high and low conditions differs only in the implicative relations between singular and dual forms. The model also included by-participant intercepts.[4]

The model revealed a significant effect of i-complexity (b=0.198, z=2.6, p<0.01), as well as a significant effect of accuracy in block 2 (b=0.83, z=10.5, p<0.001) and a significant effect of the previous trial (b=0.15, z=4.4, p<0.001), suggesting that participants were better at generalizing the dual when they had previously encountered the singular compared to the plural. The model also revealed a significant interaction between i-complexity and the previous trial (b=0.24, z=7.25, p<0.001), suggesting that the effect of i-complexity was larger for dual trials following the singular. The model did not reveal a main effect of cue systematicity (b=-0.02, z=-0.22, p=0.82) but revealed an interaction between cue systematicity and item marking (b=0.1, z=2.99, p<0.01), showing that participants in the systematic semantic cues conditions were better at generalizing the dual forms to novel stems for semantically marked items. Crucially, the model failed to reveal an interaction between i-complexity and cue systematicity (b=-0.03, z=-0.37, p=0.71) or an interaction between i-complexity, cue systematicity and item marking (b=-0.017, z=-0.5, p=0.62). Fig. 5.8 shows participants' generalization accuracy on duals by item marking and the grammatical number of the previous trial. These results suggest that systematic semantic cues for class membership did not enhance the effect of low i-complexity on generalization, although semantic cues did facilitate generalization to novel marked items.

Results from this experiment suggest that participants did notice the semantic cues when trained on languages in which these cues were systematically linked to class membership. Semantic cues, as opposed to phonological cues for class membership (chapter 4), showed a

[4]A model that included a by-participant random slope for item marking produced a singular fit warning. We therefore ran the model without this random effect

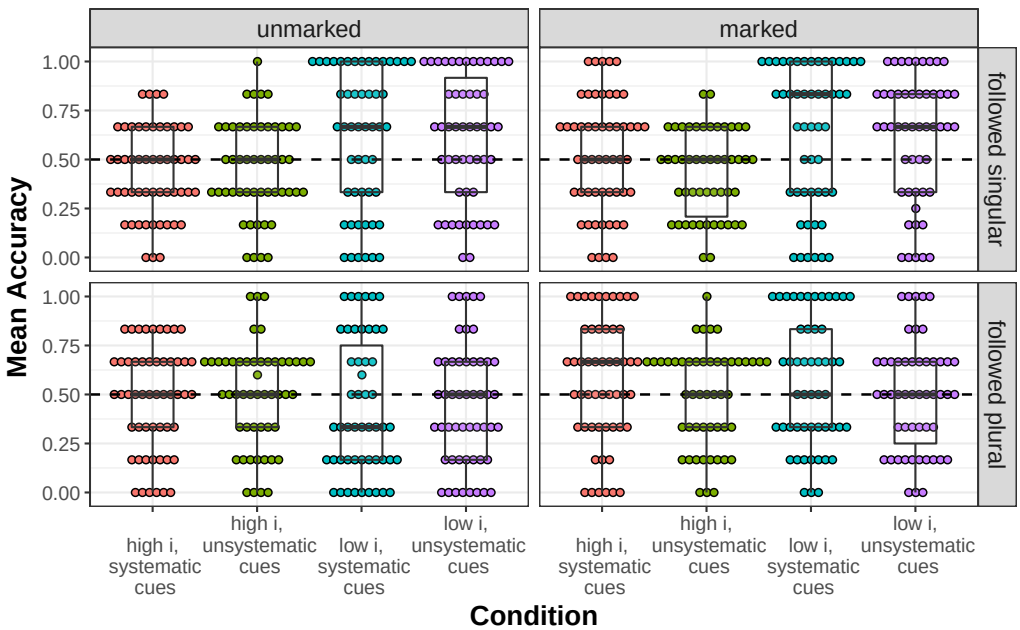


Figure 5.8: Mean accuracy for dual trials in the generalization phase split by item marking (columns) and the previous trial (rows). Dual trials that followed singular trials (upper row) show difference in performance across i-complexity conditions. Accuracy on semantically marked items (right facets) was higher than on unmarked items (left) for participants in the systematic cues conditions, suggesting for an interaction between item marking and cue systematicity in their effect on generalization to novel forms in dual.

facilitating effect both in learning the forms in the language and in generalizing the paradigm to novel stems in this task. The results also suggest that low i-complexity facilitates generalizing the paradigm to novel stems. This supports Ackerman and Malouf 2015, and provides some evidence for the advantage of low i-complexity in learning forms encountered in low frequency (though note this was seen only for better learners). These two effects, however, did not interact with each other as we hypothesised. While we found this interaction (using phonological cues) with the neural network model, learning a language with systematic semantic cues for class membership does not lead to a greater facilitative effect of low i-complexity compared to learning a language where class membership is arbitrary.

5.3 Discussion

In Chapters 4 and 5, we tested the effect of i-complexity on learning and generalizing inflectional paradigms in the presence of extra-morphological cues for class membership, both with neural networks and human participants. Previous findings with human participants suggest that generalization of morphological paradigms is enabled only in the presence of additional cues for class membership (e.g., Braine 1987; Brooks et al. 1993; Frigo and McDonald 1998; Kempe and Brooks 2001; L. A. Gerken et al. 2009), with a few exceptions showing evidence for generalization when such cues are absent (Mintz 2002; Reeder et al. 2013; Seyfarth et al. 2014). Based on these findings we hypothesised here that the effect of i-complexity (capturing distribution information) on learning and generalization would be augmented in the presence of phonological or semantic cues for class membership.

Our findings suggest that extra-morphological cues (at least when they are salient enough) facilitate both learning and generalization. Further, low i-complexity was found to facilitate generalization, and in some instances, learning - while LSTM neural networks benefited from low i-complexity, we did not see a robust effect of i-complexity on learning in human learners. Lastly, contra to our hypothesis, the addition of systematic cues to class membership did not result in stronger effects of i-complexity. A number of points can be drawn based on these findings.

First, the facilitative effect of systematic additional cues on learning and generalization was as expected, based on previous findings (e.g., Frigo and McDonald (1998), L. Gerken et al. (2005), and Ouyang et al. (2012)). Frigo and McDonald (1998) claim that the cues must be salient in order to facilitate learning and generalization. In our task, semantic cues were found to be more salient than phonological cues (i.e., facilitating learning and generalizing marked items), but we believe that this has to do with the specific task and cues we used,

rather than establishing a more general claim on the effect of systematic phonological versus semantic cues on learning and generalization.

Second, in these two chapters, we tested the effect of i-complexity on generalization in human learners in addition to its effect on learning. Low i-complexity was found to facilitate generalization to novel stems in both chapters 4 and 5. These findings are consisted with previous findings (Seyfarth et al. 2014). However, here and in Chapter 2 the effect of i-complexity on learning was found in only some of the experiments. A combined analysis of the data from the learning task in Chapter 2 and Chapter 4 suggest that the inconsistent effect is not dependent on differences in the tasks' design itself (e.g., size of the language). We concluded that the effect of i-complexity on learning in human learners is present but weak and is therefore sparsely found (e.g., in experiments with higher numbers of participants). This difference in the effect of i-complexity across tasks (learning encountered items vs. generalizing to novel items) in human learners could imply that there is a difference in the mechanisms active in these two tasks. Another interpretation of the results is that generalization serves as a 'low bar' for finding effects of predictive structure than learning encountered forms. I further discuss this difference in the General discussion.

Furthermore, the fact that i-complexity facilitates generalization in humans, regardless of the presence of additional cues, is surprising in light of previous results showing generalization only in the presence of systematic cues. It is possible that an additional distributional cue is sufficient for learners for generalizing the paradigm, even without extra-morphological cues. In our target paradigms, both the singular and the plural forms in the low i-complexity condition are predictive of the form in dual. Put another way, since the plural form is also informative of the form in dual, it can serve as an additional distributional cue on top of the form of the lexeme in singular. This suggestion is compatible with Bonami and Beniamine (2016) conjecture that language users rely on multiple known forms of a lexeme to infer a

target inflected form; they propose that guessing an unencountered inflected form can be facilitated based on knowledge of more than one other inflected form of the same lexeme. In order to assess this interpretation of the results, testing directly whether rich distribution information serves as an additional cue for paradigm generalization is required.

Third, we hypothesized that i-complexity interacts with systematic cues in their effects on learning and generalization; Frigo and McDonald (1998) show that participants in the systematic cue condition were able to produce the correct form for novel items, even when the specific novel item was not phonologically marked. Therefore, in the presence of systematic cues learners were better able to use the predictive structure of the paradigm. However, our results do not show satisfactory evidence for this interaction between systematic cues and i-complexity. In fact, we found such interaction only in generalization in LSTM neural networks. This could suggest a difference between the two types of learners, a subject we discuss in depth in the General Discussion (Chapter 6).

Chapter 6

General Discussion

In this thesis, I systematically explored the role i-complexity has in shaping natural languages through learners' inductive biases. Over the three parts of the thesis I tested the effects of i-complexity on learning and generalizing inflectional paradigms in human and neural network learners. To assess the magnitude of effects of i-complexity on morphological learning I compared these effects against effects of e-complexity. A summary of the results is presented in Table 6.1 below.

In Part 1, I tested whether learners are sensitive to i-complexity when learning inflected forms in a miniature language. First, with neural networks I replicated previous results with human learners showing an effect of i-complexity on generalizing inflectional paradigms to novel items. Second, testing the effect of i-complexity on learning trained forms with neural networks and human learners showed weak effects of i-complexity on learning; in neural networks, i-complexity was found to facilitate learning, while in human learners an effect of i-complexity was found in only one out of three experiments. Third, comparing the effect of i-complexity to the effect of e-complexity on learning, findings show evidence for greater effects of e-complexity, in both human learners and neural networks. Note that the task used in this part was designed to increase the likelihood of finding an effect of i-complexity; learners received staged training, encountering predictive forms before encountering other forms.

In Part 2, I compared the effect of i-complexity on learning with that of e-complexity and as-

sessed the relationship between these two measures, using randomly constructed paradigms. As opposed to experiments in Part 1, learners in this part were trained on forms in the paradigm in a random order; this was done to compare the effects of i- and e-complexity in a learning regime that is neutral in terms of enhancing or reducing the probability that learners would be affected by one measure or another. Again, effects of e-complexity were greater than effects of i-complexity in both learners; in neural networks, both i- and e-complexity were found to affect learning the paradigms, with low e-complexity being more advantageous. In human learners, only e-complexity was found to affect learning.

In Section 3.4, 1000 inflection paradigms were randomly generated and a strong negative correlation was revealed between i- and e-complexity. Furthermore, patterns of low i-complexity similar to the typological observations by Ackerman and Malouf (2013) were found in the random paradigms, where no inductive biases are in place. These results suggest that the typological observations may, in part, reflect an intrinsic relationship between the two measures. Finally, neural networks were trained and tested on the randomly generated paradigms and replicated previous results for varying values of i- and e-complexity.

While experiments in Part 1 and 2 were designed to eliminate extra-morphological cues for class membership, in Part 3 I tested whether the presence of phonological or semantic cues amplifies the effect of i-complexity in learning and generalizing inflectional paradigms. Results from these studies do not provide evidence for an interaction between i-complexity and extra-morphological cues on learning and generalisation. However, extra-morphological cues, when salient enough, were found to facilitate learning and generalization in both human learners and neural networks. Furthermore, low i-complexity was found to facilitate *generalizing* the paradigm to novel stems in both learners, replicating previous results.

Overall, my findings suggest that i-complexity only weakly affects learning and generalizing inflectional paradigms. Here I discuss the role i-complexity may have in shaping morpholog-

ical paradigms of natural languages, following the three exploratory themes throughout the thesis.

		i-complexity	**e-complexity**
LSTM neural networks	Staged learning regime	● ●	●
	Unstaged learning regime	●	●
Human learners	Staged learning regime	○ ○ ● ● ○	●
	Unstaged learning regime	○	●

Table 6.1: Summary table of the results. Effects of i- and e-complexity on learnability of inflection paradigms. ○ represents an experiment with no effect found while ● represents an experiment where an effect was found.

Learning and generalization

Generalizing to completely novel forms is an extreme case of a much more general problem that language learners face of producing forms which may have been encountered but have not yet been robustly acquired. I hypothesised that learners can use the same strategy they use when generalizing to completely novel stems to help generate (or recall) low frequency forms that are not fully memorized; in other words, if i-complexity facilitates generalization to novel forms, it should, in principle, facilitate learning forms under low exposure as well.

However, while results from the generalization task with human learners replicated previous findings showing effects of i-complexity on generalizing the paradigm to novel words (Seyfarth et al. 2014), results from the task of producing encountered forms (i.e., learning task) showed only weak evidence for effects of i-complexity on learning low frequency forms.

These results can be interpreted as suggesting that contra my hypothesis, the mechanisms

used in the two tasks, generalization to completely novel items and learning forms in low frequency, are different in essence; i-complexity affects the former but not the latter. However, this seems an unlikely explanation as that would presume perfect memory of which items have been encountered.

A more plausible explanation for these findings, in my eyes, is that generalization puts a low bar for finding an effect of i-complexity, since it compares a case where generalization is possible through using the predictive structure (in low i-complexity paradigms) to a case where generalization to novel forms is simply not possible (in high i-complexity paradigms). In a task of learning encountered words, however, performance higher than chance is possible in both low and high i-complexity paradigms, through memorization. Therefore, if low i-complexity facilitates the learning process, we expected to see better performance in learning word forms of low i-complexity paradigms. The inconsistency in finding an effect of i-complexity on learning encountered forms suggests that an effect of i-complexity on learning morphological paradigms may simply be present, but weak.[1]

Recall that i-complexity represents the extent to which inflectional forms in a paradigm can predict one another by analogy over the suffixes; if a word is marked with the same suffix as another word in one inflectional category, then by analogy it will be marked with the same suffix as that other word in another inflectional category (e.g., Ackerman and Malouf 2013; James P. Blevins 2006; G. T. Stump 2001). The weak effects of i-complexity on learning encountered forms may suggest that there is a difference between learners' ability to make analogies based on similarities in the stem and analogies over the suffixes. While there is evidence that phonological similarities of stems assist in learning the classification of forms (e.g., Frigo and McDonald 1998), results from this thesis suggest that analogies based on the

[1]Note as well that in the combined analyses I performed in Chapter 2 and Chapter 4, results did not show that differences in the design of the tasks (e.g., predictive trials or larger lexicons) is what led to revealing an effect of i-complexity in some of the experiments.

suffixes is more difficult for (human) learners.

Human and neural network learners

I used LSTM neural networks as a supplement to human learners in testing the relative impact of i-complexity on paradigm learning. I trained LSTM neural networks as a convenient 'ideal learner', to test whether i-complexity can in principle influence paradigm learnability, using the networks as 'subjects' in a psycholinguistic task. In Chapter 3, LSTM neural networks were used as the sole subjects in learning a large number of randomly-generated inflection paradigms with varying values of i- and e-complexity, which is less feasible with human participants.

The LSTM neural networks used here displayed learning behaviour as would be expected from an 'ideal learner'. The neural networks were sensitive to every manipulation that was set in the experiments and showed little variance across different runs of the model per task. Also when training the models with different sets of hyperparameters (hidden and embedding dimensions, learning rates and optimizers), the same patterns of results were observed for the majority of hyperparameters and no cases where the reversed patterns were exhibited (see Appendix B).

Throughout the experiments and the different tasks, results from the LSTM networks mirrored the human behaviour, to a large extent. First, in generalizing the paradigm to novel items, results from the neural networks replicated previous findings showing effects of i-complexity with human participants (Seyfarth et al. 2014) and were similar to results with human learners in Chapter 5. Second, in learning encountered forms, results from the neural networks displayed similar overall patterns to data from human learners tested on a matched task; greater effects of e-complexity were found with both learners.

Yet, there were also differences between the human and neural network learners. First, while

extra-morphological cues were not found to interact with i-complexity in human learners, not in learning nor generalizing the paradigm, in neural networks an interaction was found in generalization to novel items. Second, i-complexity was found to robustly affect learning in neural networks while only sparsely in human learners. These differences could possibly result from the fact that human learners display more differences between participants leading to 'noisier' data than data from the neural networks. However, it is also likely that neural networks employ strategies that are to some extent different than those used by human learners.

Future work is planned to tackle this last point and looking 'under the hood' of the neural networks to better understand what strategies are used during morphological systems learning and comparing them to those employed by human learners. To do so, I intend to train and test simpler feed-forward neural networks and examine the embeddings created for forms in the paradigms, for low versus high i-complexity languages. In addition to this line of work, the LSTM neural networks can be trained and tested on inflection paradigms of varying sizes, which is less feasible with human learners due to long training times; larger inflection paradigms could produce larger differences in i-complexity and can more reliably reflect paradigms of natural languages.

The role of i-complexity in language change

Ackerman and Malouf (2015) hypothesised that i-complexity shapes languages through cases of generalization to novel words; since speakers are not exposed to the full set of inflections for each lexeme in the language, they sometime have to produce inflected forms they have not yet encountered. In these cases, speakers are more likely to produce word forms which reflect predictive relationships when attempting to generalise, thus introducing errors that reduce the i-complexity of the paradigm. In this way, paradigms with low i-complexity will be relatively stable whereas paradigms with high i-complexity will tend to change. This

hypothesis is supported by results from Seyfarth et al. (2014) and from Chapters 4 and 5, showing an effect of i-complexity on generalization to novel items.

However, as discussed throughout the thesis, the case of generalizing to completely novel items is an extreme case of producing low frequency forms, for which the effect of i-complexity was less robust. Furthermore, I have pointed out that the task of generalization sets a low bar for testing the effect of i-complexity on the learnability of inflectional paradigms.

To assess the magnitude of effects of i-complexity on learnability of inflection paradigms, I compared the effects of i-complexity with those of a different measure of morphological complexity, that I refer to as e-complexity, following Ackerman and Malouf 2013. Two main aspects of this measure are (a) it accounts for complexity that originates from the number of inflection classes in the paradigm and the use of allomorphy, and (b) contrary to measures of i-complexity, this measure does not reflect the difficulty of solving the PCFP based on knowledge of other inflected forms of the same lexeme. E-complexity (measured as average cell entropy) was found to have a stronger and robust effect on learning compared to i-complexity, both in human and LSTM neural network learners.

Overall, the results from the thesis suggest that i-complexity has a weak effect on the learnability of morphological paradigms. Weak learning biases can still have a role in shaping natural languages. Bayesian models simulating cultural transmission (i.e., the process by which language is passed from person to person over generations) show that with time, languages mirror or even magnify agents' biases, even weak ones (Griffiths and Kalish 2007; Kalish et al. 2007; Kirby, Dowman, et al. 2007)[2].

Results from a set of randomly-generated paradigms (Chapter 3) suggest that e-complexity and i-complexity are strongly negatively correlated. This might suggest that inflectional

[2]Although other simulations with Bayesian learners suggest that it depends also on the communication structure; when a learner receives different inputs from more than one other language users, language may not reflect the inductive bias (K. Smith 2009)

paradigms are organized to minimize *either* i- or e-complexity. However, these findings were found in paradigms of a fixed size, and therefore should be further tested with paradigms of varying sizes.

The strong effect of e-complexity on the learnability of morphological paradigms found here suggests that the frequency of forms play an important role in the learnability of the paradigm. This is a further evidence for the pervasiveness of the effects of frequency on language learning (e.g., Ambridge et al. 2015). In the context of inflectional complexity, Sims and Parker (2016) suggest that in addition to implicative structure (i-complexity), type frequency of inflection classes also plays a role in reducing the complexity of the paradigm. Although type frequency of all noun classes was kept constant throughout the thesis, findings support the general claim that the frequency of elements in the paradigm plays a role in inferring the correct inflected form for a lexeme.

Since e-complexity was found to be a main predictor of the learnability of inflection paradigms, paradigms with low e-complexity should be more dominant cross-linguistically, all other things being equal. However, typological observations suggest that e-complexity in natural languages vary significantly (Ackerman and Malouf 2013). Other frequency effects may also influence the learnability of inflection paradigms (e.g., inflection classes type/token frequencies or frequencies of forms of grammatical functions in the paradigm). Potentially, in high e-complexity paradigms these additional frequency effects might play a role in reducing learning-relevant complexity. Future research should explore the relationship between different frequency effects both typologically and with respect to learnability.

To summarize, findings from the thesis suggest that a number of factors affect the learnability of inflection paradigms, to a different degree. Moreover, as previous studies show interdependence between measures of morphological complexity (e.g., Bane 2008; Bentz et al. 2016; Shosted 2006), results from this thesis show that i-complexity correlates as well with other

features of the paradigm examined here (specifically e-complexity and number of unique affixes). Overall, these findings suggest that in order to study learnability of morphological systems, different types of complexity should be explored jointly.

Conclusions

Ackerman and Malouf (2013) postulated the Low Conditional Entropy Conjecture based on their typological observations; they suggest that predictive structure is a shared feature of large inflection paradigms of natural languages. In a series of artificial language learning tasks both with human learners and LSTM neural networks, I tested the hypothesis that predictive structure (measured using i-complexity) influences the learnability of inflection paradigms.

Results show weak evidence for an effect of i-complexity on learning, with evidence for greater effects of e-complexity in both human and neural network learners. A strong negative correlation was found between i-complexity and e-complexity, suggesting that paradigms with higher surface complexity tend to have more predictive structure, as measured by i-complexity. There is no evidence for an interaction between i-complexity and extra-morphological cues on learning and generalisation. This suggests that semantic or phonological cues for class membership, which are common in natural languages, do not enhance the effect of i-complexity on learning and generalisation. Finally, i-complexity was found to affect generalisation in both human and neural network learners, suggesting that i-complexity could, in principle, shape languages through the process of generalisation to unknown forms.

Although it may well be that learners use predictive structure when generalizing to completely novel forms, findings from comparing effects of i-complexity on learning encountered forms suggest that pressure from learning should tend to favour low e-complexity rather than

low i-complexity.

Appendices

Appendix A

Appendix for Part I

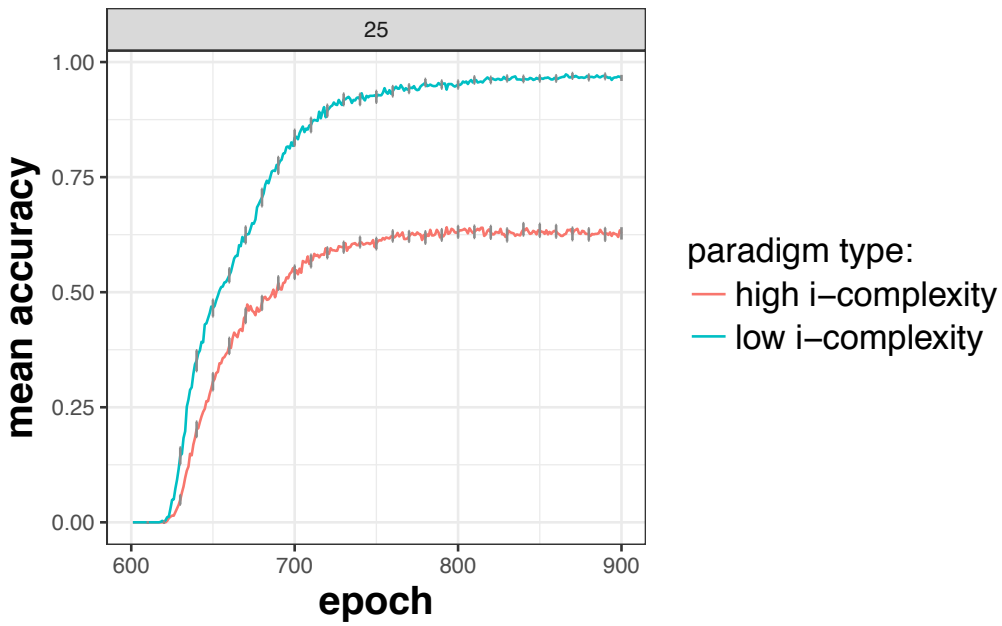


Figure A.1: Accuracy of the LSTM networks in generalizing to dual forms of novel stems, after trained on their forms in singular only, for the low i-complexity paradigm (blue) and the high i-complexity paradigm (red). for one network of size 25, with error bars indicating standard error every 10 epochs. Note that the plots start at epoch 600, when the dual forms are introduced to the network (at the beginning of Block 3). Networks trained on the high i-complexity paradigm reach accuracy of around 66% which is the expected chance accuracy when guessing according to the frequent suffix for dual, while networks trained on the low i-complexity paradigm still show accuracy of almost 100%.

Appendix B

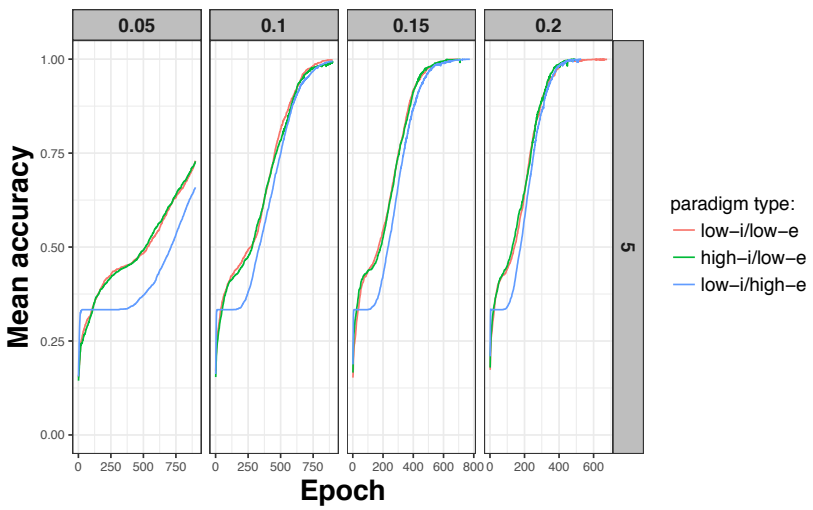
Appendix for Part II

Exploring hyperparameters space

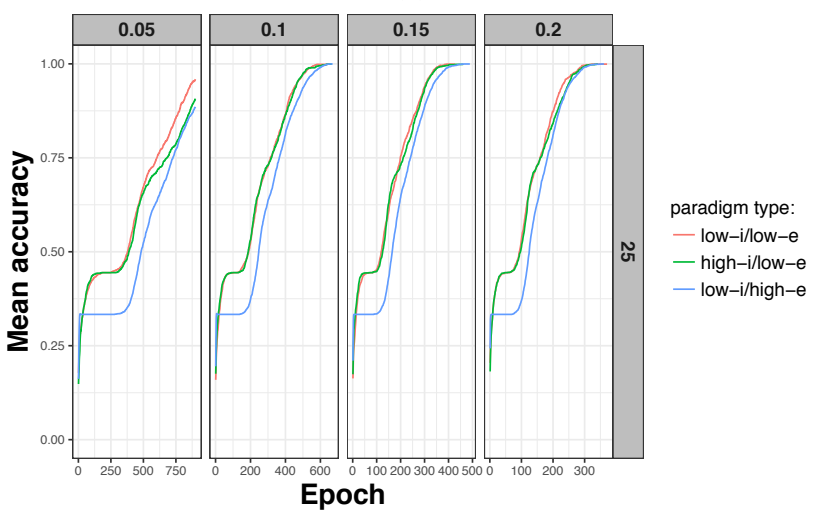
For the LSTM model presented in Section 3.3.2 we explored further hyperparameters in addition to the parameter settings specified in the main text. We explored two optimizers, SGD and Adam (Kingma and Ba 2014). We used these two optimizers with networks of of two hidden and embedding dimensions (5 and 25), trained with four different learning rates. Since we were interested in the cases were the networks fully learned the forms in the language by the end of 900 epochs, the explored learning rates differed across optimizers; for models optimized with SGD, we explored learning rates of 0.05, 0.1, 0.15 and 0.2. For models optimized with Adam, where learning was more rapid, we explored learning rates of 0.0005, 0.001, 0.0015 and 0.002.

Results are presented in Figures B.1-B.4, and a summary of the mean summed accuracy for all combinations of hyperparameters is presented in Tables B.1, B.2 below. Results from all models optimized with SGD show small effects of i-complexity compared to effects of e-complexity, regardless of the learning rate of the network. Models optimized with Adam show a similar trend for the very low learning rates, but for the rest of the models there is no difference between the conditions. Crucially, none of the hyperparameters combinations we explored showed the opposite picture where i-complexity has a stronger effect on learning than e-complexity.

These results show that for this space of hyperparameters, all models replicate the results presented in Section 3.3.2, namely that in cases where i-complexity has an effect on learning the paradigm, the effect is smaller than the effect of e-complexity.



(a)



(b)

Figure B.1: Learning trajectories of networks with two embedding and hidden layer dimensionalities; (a) networks with 5-dimensional embedding vectors and hidden layer, (b) networks with 25-dimentional embedding vectors and hidden layer, trained with different learning rates (columns), and optimized with SGD. X axis shows number of epochs up to perfect learning of the forms in the language (differs across learning rates and network dimensions).

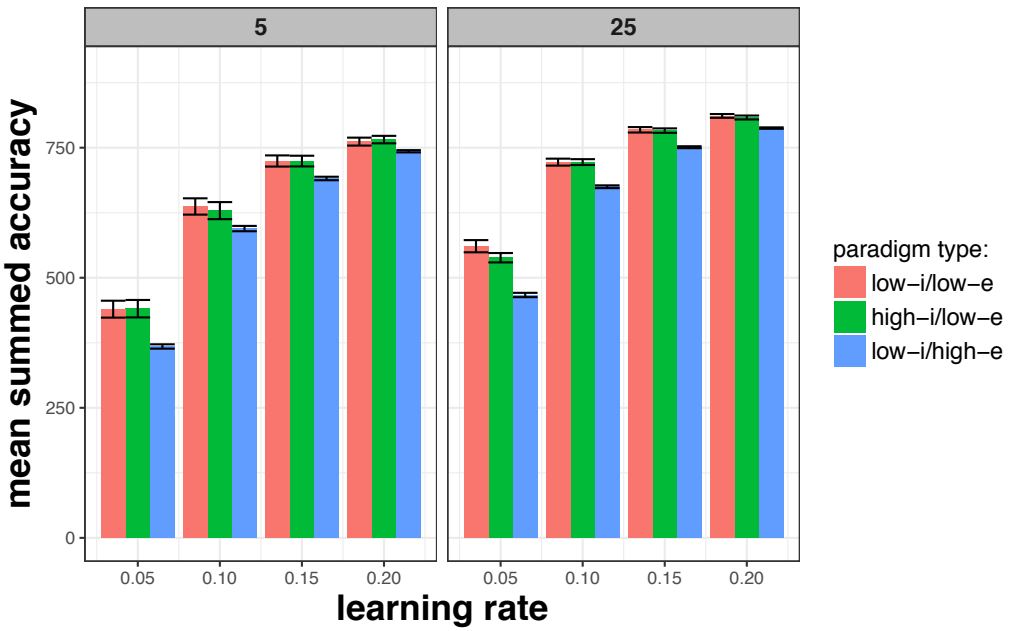
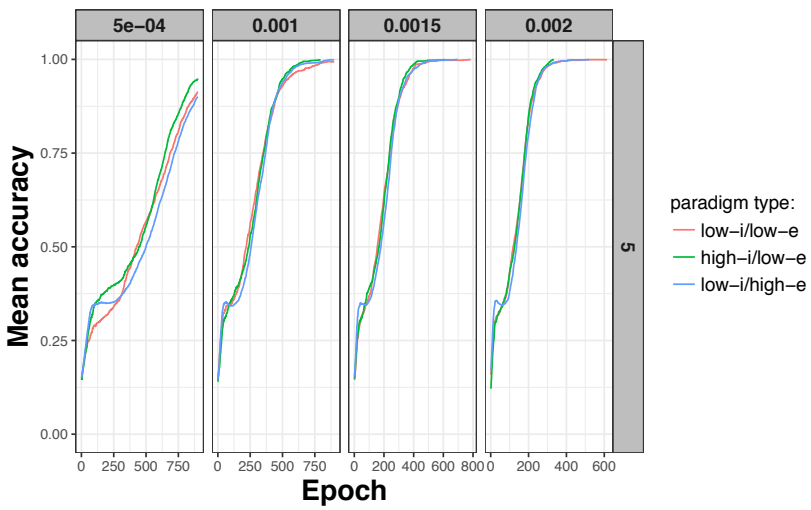


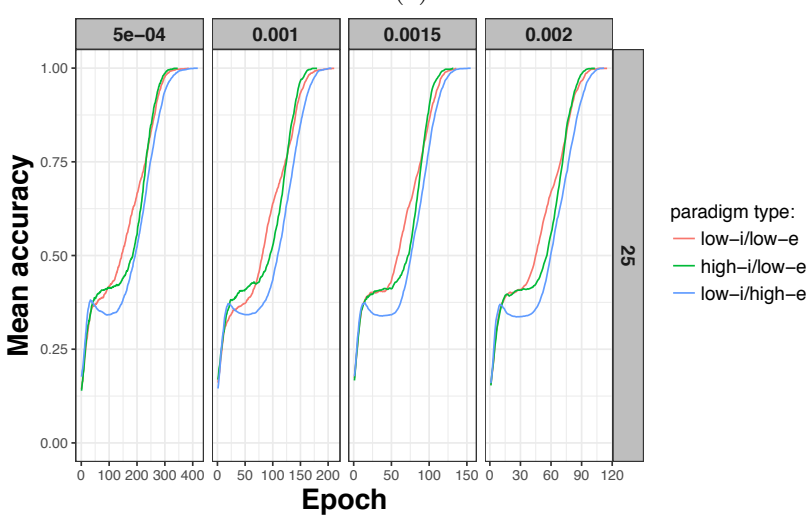
Figure B.2: Summed accuracy over the 900 epochs of the networks trained on each of the three paradigm types for models with different learning rates (x axis) and for models with different dimensions (columns) optimized with SGD.

		5				25			
		0.05	**0.1**	**0.15**	**0.2**	**0.05**	**0.1**	**0.15**	**0.2**
SGD	low-i	439.6	637.0	724.4	761.7	560.6	722.2	784.4	811.1
	/low-e	(48.7)	(47.0)	(32.2)	(22.9)	(35.1)	(20.0)	(15.6)	(10.82)
	low-i	440.5	629.0	724.3	765.6	538.5	722.3	782.9	808.0
	/low-e	(50.3)	(49.2)	(30.8)	(21.6)	(27.1)	(16.9)	(12.7)	(11.4)
	low-i	367.9	594.5	690.8	743.1	466.8	674.9	750.9	787.7
	/low-e	(41.4)	(51.0)	(33.5)	(21.4)	(41.4)	(26.5)	(18.1)	(13.4)

Table B.1: Summary of mean of summed accuracy of the model runs optimized with SGD with combinations of hidden and embedding dimensions (5, 25) and learning rates (0.05, 0.1, 0.15, 0.2). Standard deviations are presented in brackets.



(a)



(b)

Figure B.3: Learning trajectories of networks with two embedding and hidden layer dimensionalities; (A) networks with 5-dimensional embedding vectors and hidden layer, (b) networks with 25-dimentional embedding vectors and hidden layer, trained with different learning rates (columns), and optimized with Adam (Kingma and Ba 2014). X axis shows number of epochs up to perfect learning of the forms in the language (differs across learning rates and networks dimensions).

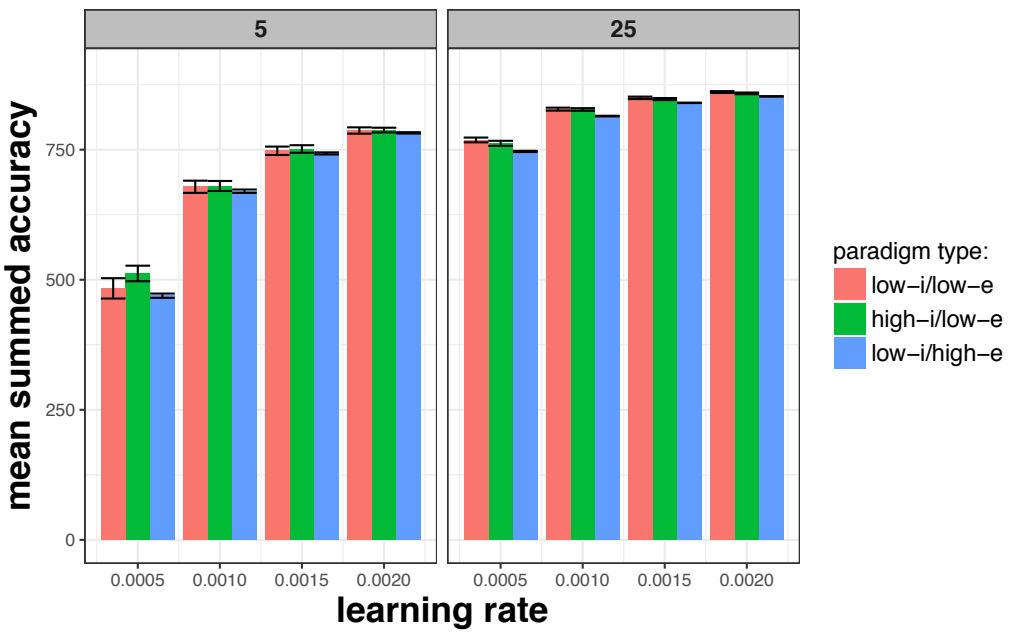


Figure B.4: Summed accuracy over the 900 epochs of the networks trained on each of the three paradigm types for models with different learning rates (x axis) and for models with different dimensions (columns) optimized with Adam.

		5				**25**			
		0.0005	**0.001**	**0.0015**	**0.002**	**0.0005**	**0.001**	**0.0015**	**0.002**
Adam	low-i	483.5	678.7	747.9	786.9	786.7	827.9	849.7	860.8
	/low-e	(58.7)	(35.2)	(24.2)	(18.3)	(13.8)	(8.5)	(7.1)	(5.2)
	low-i	512.1	680.3	751.4	787.7	762.2	827.3	847.3	858.2
	/low-e	(44.8)	(28.8)	(21.7)	(13.5)	(14.6)	(7.5)	(5.9)	(4.9)
	low-i	469.3	670.2	742.9	782.3	746.6	814.6	840.1	852.5
	/low-e	(40.9)	(32.0)	(20.11)	(13.0)	(11.4)	(5.9)	(3.8)	(3.3)

Table B.2: Summary of mean of summed accuracy of the model runs optimized with Adam with combinations of hidden and embedding dimensions (5, 25) and learning rates (0.0005, 0.001, 0.0015, 0.002). Standard deviations are presented in brackets.

Bibliography

Ackerman, F., Blevins, J. P. [James P.], & Malouf, R. (2009). Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar: Form and acquisition* (pp. 54–82). Oxford University Press.

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, *89*(3), 429–464.

Ackerman, F., & Malouf, R. (2015). The no blur principle effects as an emergent property of language systems. *Proceedings of the annual meeting of the Berkeley Linguistics Society*, *41*, 1–14.

Aĭkhenval d, A. Y. (2000). *Classifiers : A typology of noun categorization devices*. Oxford University Press.

Ambridge, B. (2010). Children's judgments of regular and irregular novel past-tense forms: New data on the english past-tense debate. *Developmental Psychology*, *46*(6), 1497.

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, *40*(5-6), 509–559.

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*(2), 239–273.

Aronoff, M. (1994). *Morphology by itself: Stems and inflectional classes*. MIT Press.

Baerman, M., Brown, D., & Corbett, G. G. (2005). *The syntax-morphology interface: A study of syncretism*. Cambridge University Press.

Baerman, M., Brown, D., & Corbett, G. G. (2010). *Morphological complexity: A typological perspective* [Unpublished manuscript, University of Surrey]. https://www.researchgate.net/publication/266215146

Baerman, M., Brown, D., & Corbett, G. G. (2015). *Understanding and measuring morphological complexity*. Oxford University Press.

Bane, M. (2008). Quantifying and measuring morphological complexity. *Proceedings of the 26th west coast conference on formal linguistics*, 69–76.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the mind: An introduction to parallel processing in networks*. Basil Blackwell.

Bentz, C., Ruzsics, T., Koplenig, A., & Samardzic, T. (2016). A comparison between morphological complexity measures: Typological data vs. language corpora. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 142–153.

Bickel, B., & Nichols, J. (2007). Inflectional morphology. *Language typology and syntactic description* (pp. 169–240). Cambridge University Press.

Bickel, B., & Nichols, J. (2013). Inflectional synthesis of the verb. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/22

Blevins, J. P. [James P.]. (2006). Word-based morphology. *Journal of Linguistics*, *42*(3), 531–573.

Blevins, J. P. [James P], Ackerman, F., Malouf, R., & Ramscar, M. (2016). Morphology as an adaptive discriminative system. *Morphological metatheory*, 271–302.

Bonami, O., & Beniamine, S. (2016). Joint predictiveness in inflectional paradigms. *Word Structure*, *9*(2), 156–182.

Braine, M. D. (1987). What is learned in acquiring word classes—a step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of language acquisition: The 20th annual carnegie mellon symposium on cognition*. Routledge.

Brodsky, P., & Waterfall, H. (2007). Characterizing motherese: On the computational structure of child-directed language. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *29*(29).

Brooks, P. J., Braine, M. D., Catalano, L., Brody, R. E., & Sudhalter, V. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of memory and language*, *32*(1), 76–95.

Bybee, J. L. (1995). Diachronic and typological properties of morphology and their implications for representation. In L. B. Feldman (Ed.), *Morphological aspects of language processing* (pp. 46–225). L. Erlbaum Associates.

Canini, K. R., Griffiths, T. L., Vanpaemel, W., & Kalish, M. L. (2014). Revealing human inductive biases for category learning by simulating cultural transmission. *Psychonomic Bulletin & Review*, *21*(3), 785–793.

Chaitin, G. J. (1988). *Algorithmic information theory*. Cambridge University Press.

Chater, N., Clark, A., Goldsmith, J. A., & Perfors, A. (2015). *Empiricism and language learnability*. Oxford University Press.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in cognitive sciences*, *7*(1), 19–22.

Chollet, F. et al. (2015). Keras.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *The Behavioral and Brain Sciences*, *31*(5), 489–509.

Corbett, G. G. [Greville G.]. (1991). *Gender*. Cambridge University Press.

Corbett, G. G. [Greville G.]. (2005). The number of genders (chapter and map). *The world atlas of language structures* (pp. 126–137). University of Surrey.

Corbett, G. G. [Greville G.]. (2009). Suppletion: Typology, markedness, complexity. In P. O. Steinkrüger & M. Krifka (Eds.), *On inflection* (p. 40). Mouton de Gruyter.

Corbett, G. G. [Greville G.]. (2013). Systems of gender assignment. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/32

Corkery, M., Matusevych, Y., & Goldwater, S. (2019). Are we there yet? encoder-decoder neural networks as cognitive models of english past tense inflection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3868–3877.

Cotterell, R., Kirov, C., Hulden, M., & Eisner, J. (2019). On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, *7*, 327–342.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., & Hulden, M. (2016). The sigmorphon 2016 shared task—morphological reinflection. *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 10–22.

Covington, M. A., & McFall, J. D. (2010). Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, *17*(2), 94–100.

Culbertson, J., Gagliardi, A., & Smith, K. (2017). Competition between phonological and semantic cues in noun class learning. *Journal of Memory and Language*, *92*, 343–358.

Culbertson, J., Jarvinen, H., Haggarty, F., & Smith, K. (2019). Children's sensitivity to phonological and semantic cues during noun class learning: Evidence for a phonological bias. *Language*, *95*(2), 268–293.

Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in psychology*, *6*, 1964.

Culbertson, J., & Newport, E. L. [Elissa L.]. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, *139*(6), 71–82.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329.

Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the brain*. Allen Lane the Penguin Press.

del Prado Martın, F. M., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, *94*(1), 1–18.

Dixon, R. M. (1986). Noun classes and noun classification in typological perspective. In C. Craig (Ed.), *Noun classes and categorization: Proceedings of a symposium on categorization and noun classification* (pp. 105–112). John Benjamins Amsterdam.

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2017). A functional theory of gender paradigms. In F. Kiefer, J. P. Blevins, & H. Bartos (Eds.), *Perspectives on morphological organization* (pp. 212–239). Brill.

Elman, J. L. [Jeffrey L.]. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L. [Jeffrey L.]. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*(2), 195–225.

Elman, J. L. [Jeffrey L], Bates, E. A., & Johnson, M. H. (1996). *Rethinking innateness: A connectionist perspective on development* (Vol. 10). MIT Press.

Elsner, M., Sims, A. D., Erdmann, A., Hernandez, A., Jaffe, E., Jin, L., Johnson, M. B., Karim, S., King, D. L., Lamberti Nunes, L., Oh, B.-D., Rasmussen, N., Shain, C., Antetomaso, S., Dickinson, K. V., Diewald, N., McKenzie, M., & Stevens-Guille, S. (2019). Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute? *Journal of Language Modelling*, 53–98.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*(44), 17897–17902.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current directions in psychological science : a journal of the American Psychological Society*, *12*(6), 227–232.

Feldman, J. (2016). The simplicity principle in perception and cognition: The simplicity principle. *Wiley interdisciplinary reviews. Cognitive science*, *7*(5), 330–340.

Finkel, R., & Stump, G. (2007). Principal parts and morphological typology. *Morphology*, *17*(1), 39–75.

Fraser, N., & Corbett, G. G. [Greville G]. (2000). Gender assignment: A typology and a model. In G. Senft (Ed.), *Systems of nominal classification (language, culture and cognition 4)* (pp. 293–325). Cambridge University Press.

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language*, *39*(2), 218–245.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 32–42.

Gerken, L. A., Wilson, R., Gómez, R., & Nurmsoo, E. (2009). The relation between linguistic analogies and lexical categories. In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar: Form and acquisition*. Oxford University Press.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of child language*, *32*(2), 249.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational linguistics*, *27*(2), 153–198.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive science*, *31*(3), 441–480.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 1195–1205.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hockett, C. F. (1954). Two models of grammatical description. *Word*, *10*(2-3), 210–234.

Hudson Kam, C. L. [Carla L.], & Newport, E. L. [Elissa L.]. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195.

Hudson Kam, C. L. [Carla L], & Newport, E. L. [Elissa L]. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66.

Johnson, T., Culbertson, J., Rabagliati, H., & Smith, K. (2020). *Assessing integrative complexity as a predictor of morphological learning using neural networks and artificial language learning* [Unpublished manuscript, University of Edinburgh]. https://psyarxiv.com/yngw9/

Johnson, T., Gao, K., Smith, K., Rabagliati, H., & Culbertson, J. (2021). Predictive structure or paradigm size? investigating the effects of i-complexity and e-complexity on the learnability of morphological systems. *Journal of Language Modelling*, *9*(1).

Jordan, M. I. (1997). Serial order: A parallel distributed processing approach. In J. W. Donahoe & V. Packard Dorsel (Eds.), *Neural-network models of cognition* (pp. 471–495). Elsevier.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, *5*(3), 206–213.

Juola, P. (2008). Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), *Language complexity: Typology, contact, change* (pp. 89–109). John Benjamins Publishing Company Amsterdam.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kempe, V., & Brooks, P. J. (2001). The role of diminutives in the acquisition of russian gender: Can elements of child-directed speech aid in learning morphology? *Language Learning*, *51*(2), 221–256.

Kettunen, K. (2014). Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, *21*(3), 223–245.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition* (pp. 173–204). Cambridge University Press.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241–5245.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Kirov, C., & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, *6*, 651–665.

Kolmogorov, A. N. (1968). Three approaches to the quantitative definition of information. *International journal of computer mathematics*, *2*(1-4), 157–168.

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. et al. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Lakoff, G. (1987). *Women, fire, and dangerous things : What categories reveal about the mind*. University of Chicago Press.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

Maldonado, M., & Culbertson, J. (2019). Something about "us": Learning first person pronoun systems. *Proceedings of the 41st annual meeting of the Cognitive Science Society*, 749–755.

Malouf, R. (2017). Abstractive morphological learning with a recurrent neural network. *Morphology*, *27*(4), 431–458.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development*. Springer.

Marzi, C., Ferro, M., Nahli, O., Belik, P., Bompolas, S., & Pirrelli, V. (2018). Evaluating inflectional complexity crosslinguistically: A processing perspective. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological science*, *2*(6), 387–395.

McCurdy, K., Goldwater, S., & Lopez, A. (2020). Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for german plurals. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1745–1756.

McWhorter, J. H. (2001). The worlds simplest grammars are creole grammars. *Linguistic typology*, *5*, 125–166.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

Meinhardt, E., Malouf, R., & Ackerman, F. (2019). *Morphology gets more and more complex, unless it doesn't* [Unpublished manuscript, San Diego State University and University of California San Diego]. https://www.researchgate.net/publication/333194657

Milin, P., Đurđević, D. F., & del Prado Martín, F. M. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from serbian. *Journal of Memory and Language*, *60*(1), 50–64.

Milin, P., Keuleers, E., & Đurđević, D. (2011). Allomorphic responses in serbian pseudonouns as a result of analogical learning. *Acta Linguistica Hungarica*, *58*(1), 65–84.

Mintz, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition*, *30*(5), 678–686.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117.

Mirman, D. (2017). *Growth curve analysis and visualization using r* (First edition.). CRC Press.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*(2), 143–182.

Moreton, E., & Pater, J. (2012). Structure and substance in artificial-phonology learning, part i: Structure. *Language and Linguistics Compass*, *6*(11), 686–701.

Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., & Kirby, S. (2019). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition*, *192*, 103964–103964.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: learning, memory, and cognition*, *14*(1), 54.

Ouyang, L., Boroditsky, L., & Frank, M. (2012). Semantic coherence facilitates distributional learning of word meanings. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*(34).

Pertsova, K. (2012). Logical complexity in morphological learning: Effects of structure and null/overt affixation on learning paradigms. *Annual meeting of the Berkeley Linguistics Society*, *38*, 401–413.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73–193.

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive science*, *26*(3), 303–343.

Ralli, A. (2002). The role of morphology in gender determination: Evidence from modern greek. *Linguistics*, *40*(3), 519–551.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive psychology*, *66*(1), 30–54.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 237–263). John Wiley & Sons, Inc.

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Transactions on Information theory*, *30*(4), 629–636.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs. *Parallel distributed processing*. The MIT Press.

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental psychology*, *39*(3), 484.

Sampson, G., Gil, D., & Trudgill, P. (2009). *Language complexity as an evolving variable* (Vol. 13). Oxford University Press.

Sapir, E. (2012). *Language : An introduction to the study of speech*. Andrews UK Limited.

Senft, G. (2000). *Systems of nominal classification*. Cambridge University Press.

Seyfarth, S., Ackerman, F., & Malouf, R. (2014). Implicative organization facilitates morphological learning. *Annual meeting of the Berkeley Linguistics Society*, *40*, 480–494.

Shannon, C. E. (1963). *The mathematical theory of communication*. University of Illinois Press.

Shosted, R. K. (2006). Correlating complexity: A typological approach. *Linguistic Typology*, *10*(1), 1–40.

Silvey, C., Kirby, S., & Smith, K. (2015). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*, *39*(1), 212–226.

Sims, A. D., & Parker, J. (2016). How inflection class systems work: On the informativity of implicative structure. *Word Structure*, *9*(2), 215–239.

Smith, E. E., & Medin, D. L. (2013). *Categories and concepts*. Harvard University Press.

Smith, K. (2009). Iterated learning in populations of bayesian agents. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *31*(31).

Stump, G., & Finkel, R. A. (2013). *Morphological typology: From word to paradigm* (Vol. 138). Cambridge University Press.

Stump, G. T. (2001). *Inflectional morphology: A theory of paradigm structure* (Vol. 93). Cambridge University Press.

Tal, S., & Arnon, I. (2018). Ses effects on the use of variation sets in child-directed speech. *Journal of child language*, *45*(6), 1423–1438.

Waterfall, H. R. (2005). *A little change is a good thing: Feature theory, language acquisition and variation sets* (Doctoral dissertation). Chicago University.

Wilmoth, S., & Mansfield, J. (2021). Inflectional predictability and prosodic morphology in pitjantjatjara and yankunytjatjara. *Morphology*, 1–27.

Wonnacott, E., & Newport, E. L. [Elissa L.]. (2005). Novelty and regularization: The effect of novel instances on rule formation. *BUCLD 29: Proceedings of the 29th annual Boston University conference on language development*, 663–673.

Xanthos, A., & Gillis, S. (2010). Quantifying the development of inflectional diversity. *First language*, *30*(2), 175–198.

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, *40*(8), 2081–2094.

Zaslavsky, N., Kemp, C., Tishby, N., & Regier, T. (2020). Communicative need in colour naming. *Cognitive Neuropsychology*, *37*(5-6), 312–324.