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

Tamar Johnson[1], *Kexin Gao*[1], *Kenny Smith*[1], *Hugh Rabagliati*[2], *and Jennifer Culbertson*[1]
[1] Centre for Language Evolution, University of Edinburgh
[2] Department of Psychology, University of Edinburgh

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 a recent work by Johnson *et al.* (2020) suggesting that both neural networks and human learners may

Keywords: morphological complexity, learning, neural networks, typology

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 and Malouf's hypothesis, as well as the role of ease of learning versus generalization to novel forms in the evolution of paradigms.

1 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. Shosted 2006; Bickel and Nichols 2013). In addition to the 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; Corbett 2009), which has also been claimed to be a source of complexity in morphological systems (e.g. Baerman *et al.* 2010; Ackerman and Malouf 2013). These aspects of morphological complexity, which pertain to the size of a morphological sys-

tem, 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. Blevins 2006; Ackerman and Malouf 2013). 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, 2015). Predictive structure in a morphological system can be seen in Table 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

Table 1: Russian nominal inflection paradigm (phonological transcription). 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, DAT = dative, LOC = locative, INS = instrumental)

	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

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 morphological 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. Deacon 1997; Kirby 2002; Christiansen and Chater 2008; Kirby *et al.* 2008; Culbertson and Kirby 2016). 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.

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; Pertsova 2012; Maldonado and Culbertson 2019). For example, in Table 1, *-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 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 provide proof-in-principle for whether i-complexity can affect learnability and whether its 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.

Measuring i-complexity and e-complexity 1.1

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 (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 in Table 1, 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.

$$(1) \qquad H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} P(x, y) \log_2 P(y|x)$$

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

$$(2) \qquad \frac{\sum_{Y \in G} \sum_{X \in G} H(X|Y)}{N_G(N_G - 1)},$$

where G is the set of grammatical functions in the paradigm and N_G is their total number.[1]

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) 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

[1] 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.

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 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.[2] 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.

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

[2] 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 (p.c.) has suggested that the 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 the number of forms – an uncontroversial measure of e-complexity – can also be considered predictive in this way, as 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.

E-complexity is measured as the averaged cell entropy over all grammatical functions in a paradigm as in (4),

$$(4) \qquad \frac{\sum_{X \in G} H(X)}{N_G},$$

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

1.2 *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.[3] 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 *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

[3] The relationship between e-complexity and i-complexity found by Ackerman and Malouf (2013) is also reported in Cotterell *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) .

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 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 the Ackerman *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 2a). In the test phase, they asked participants to generate inflected forms for a novel lexeme given that lexeme's 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 2a, for example, being prompted with a novel singular form marked with *-yez* allows the learner to predict what form the lexeme

Table 2:
Artificially constructed nominal inflection paradigms used in Seyfarth *et al.* (2014)

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

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

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

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

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 2b). 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 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 *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.

2 TESTING THE EFFECTS OF E- AND I-COMPLEXITY IN HUMAN LEARNERS AND LSTM NEURAL NETWORKS

Johnson *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 *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 *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. Wonnacott and Newport 2005; Hudson Kam and

Newport 2009; Moreton and Pater 2012; Fedzechkina *et al.* 2012, and many others). Moreover, studies using artificial learning paradigms show correspondence between lab-based learning biases and typology (e.g. see for reviews Culbertson *et al.* 2012; Culbertson and Newport 2015).

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. learn-

ing of very large paradigms or paradigms inflecting over very large lexicons.

2.1 *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 *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 2.3 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 below but with a different mapping of affixes to cells for each human participant.

As summarized in Table 3, the paradigms differ either in i-complexity or e-complexity, holding the other constant. We also hold constant all other aspects of the paradigms: the paradigms are matched in

Table 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)

		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 4: Example paradigms for each type tested. See Table 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})

(a) low-i/low-e

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

(b) high-i/low-e

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

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

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

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

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

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 4a), predicts the dual affix (e.g. *-um*). This is not the case in the high-i/low-e paradigm (in Table 4b the singular *-op* does not uniquely determine the form of the dual). The remaining two paradigms (Table 4c, d) 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 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 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 4c, d). In the low-i/high-e_{within} the syncretic form is reused as a plural in class 1 (Table 4c). In the low-i/high-e_{across} the syncretic form is reused as a plural marker for class 2 (Table 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; Pertsova 2012; Maldonado and Culbertson 2019), 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 *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.

2.2 *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 (Elman 1990; Jordan 1997; Elman 1991). 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.

2.2.1 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. An overview of the network structure is given in Figure 1. 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.[4] Initial weights were randomly generated, according to a

[4] 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.

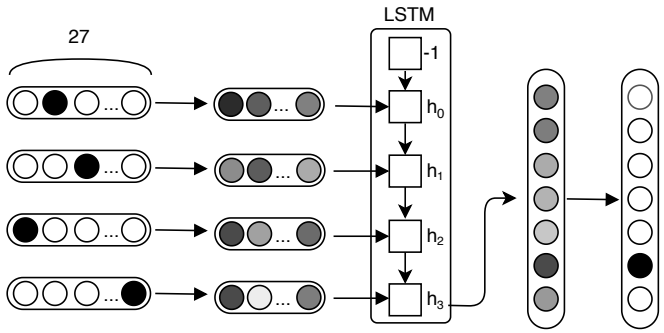


Figure 1: A diagram of the recurrent neural network: the input layer receives a string of four characters (stem + grammatical number), each coded as a one-hot vector of the length of the different characters used in the language (27). The input vectors are embedded and the embeddings are transferred 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

'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 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.[5]

Results 2.2.2

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

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

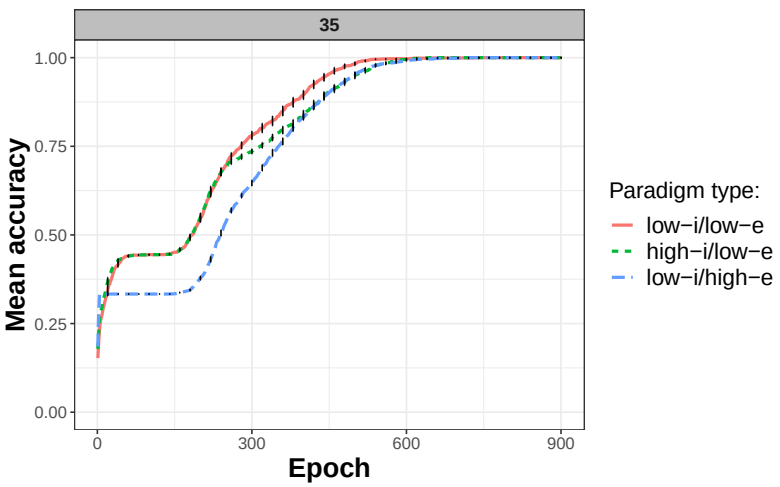
network size). For simplicity, we first collapse the two low-i/high-e paradigms in these graphs, and deal with the effect of syncretism separately below. Figure 2 presents network learning trajectories for these three paradigm types.

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. [6]

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 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 the curve we see in Figure 2, and is still interpretable for network parameterisations that do not result in convergence to 100% accuracy.

[6] 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)

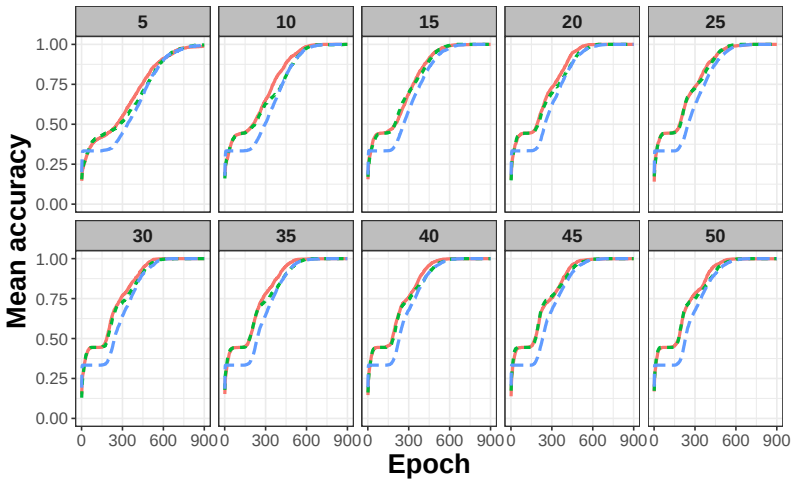


Figure 2: Network learning trajectories. (a) results for one network size (35 cells), with error bars indicating standard error every 10 epochs, (b) results for all network sizes tested (facet titles give network size in number of cells). Networks trained on low-i/low-e and high-i/low-e paradigm types 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 the Appendix for detailed results)

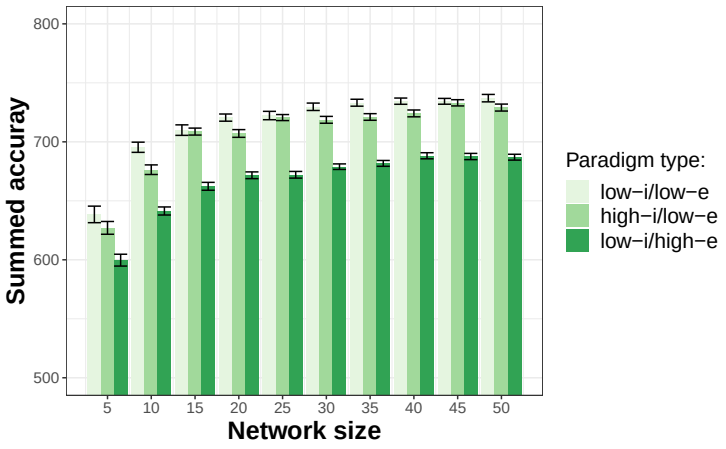


Figure 3: Summed accuracy over the 900 epochs of the networks trained on each of the three paradigm types across different sizes of the network. Error bars represent standard error. Note that the two low-i/high-e paradigms are collapsed here

Figure 3 shows the summed accuracy of the networks trained on each paradigm type across different network sizes. To determine whether these differences between network learning trajectories are significant, we ran a linear mixed-effect regression model[7] predicting 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 *et al.* (2020) we predict low-i/low-e to be the 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 ($\beta = -10.61$, $sd = 0.52$, $t = -20.23$, $p < 0.001$)

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

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 2 and in Figure 3.

2.2.3 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 4). In general, cross-class syncretism can affect both i-complexity and e-complexity, but for our paradigms neither i-complexity nor e-complexity distinguish between syncretism types; the two paradigm types have the same values for both measures. Figure 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. Pertsova 2012; Maldonado and Culbertson 2019).

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 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, the model included 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 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.

Since the type of syncretism was found to affect learning, we conducted an additional analysis to determine whether the effect of

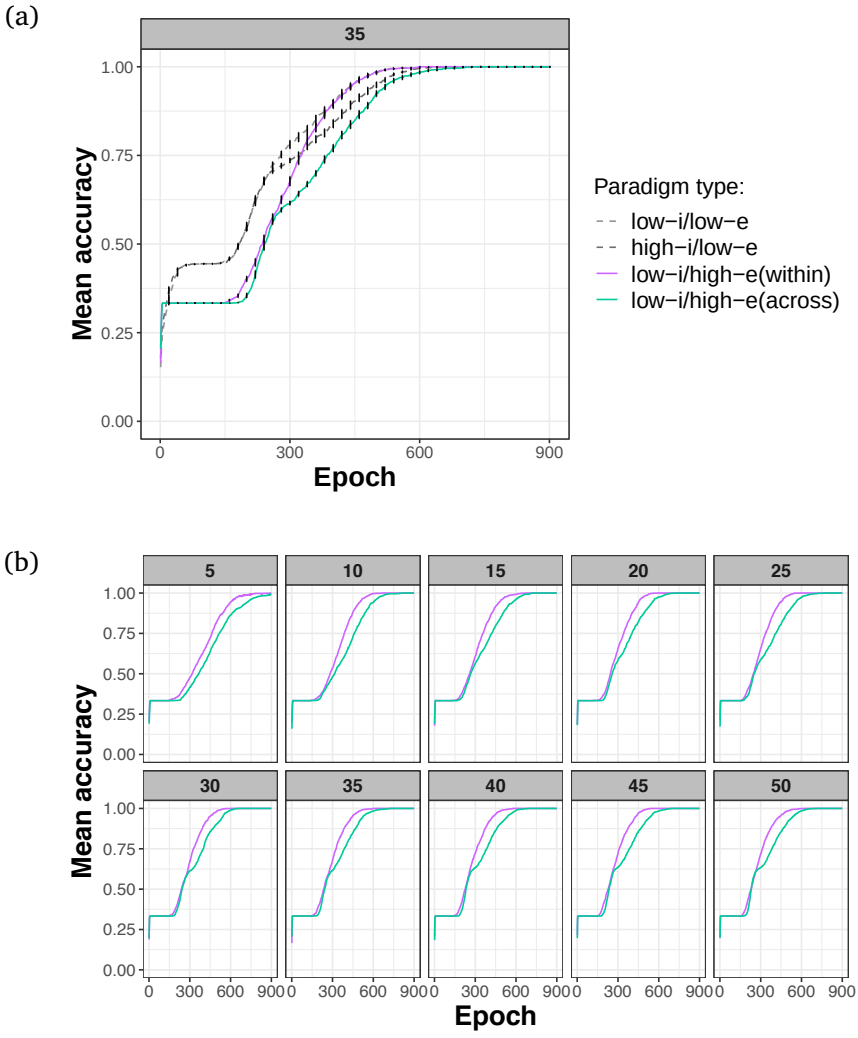


Figure 4: Network learning trajectories with low-i/high-e_{within} and low-i/high-e_{across} paradigms plotted separately. Trajectories for networks trained on low-i/low-e and high-i/low-e paradigms presented in grey (dashed lines) for comparison. (a) results for one network size (35 cells), with error 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

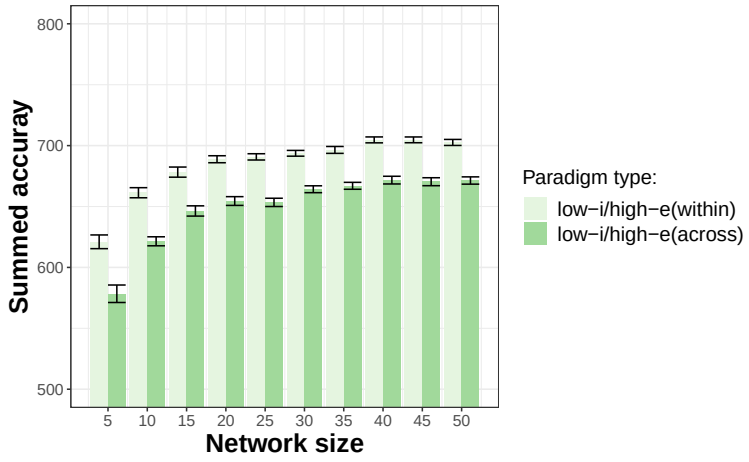


Figure 5:
Summed accuracy over the 900 epochs of networks trained on low-i/high-e_{within} and low-i/high-e_{across} paradigms across different network sizes. Error bars represent standard error. Across all network sizes the paradigm type with across-class syncretism is learned slower

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 complements earlier results for human learners and LSTMs (Seyfarth *et al.* 2014; Johnson *et al.* 2020) 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 *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.

2.3 *Experiment 2: human learners*

2.3.1 Materials

The same artificially constructed paradigms described in Table 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 *et al.* (2020). Singular nouns corresponded to a single object, dual corresponded to two objects, and plural ranged from 3 to 12 objects (selected randomly). See Figure 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.

2.3.2 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,[8] thus the final dataset consisted of 120 participants: low-i/low-e (29); high-i/low-e (31); low-i/high-e_{within} (28); low- i/high-e_{across} (31).

2.3.3 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 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.

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

Figure 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

Results 2.3.4

Figure 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 7). Therefore, we used logistic growth curve analysis (Mirman 2017) to analyse the effect of i-complexity and e-complexity on learning 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

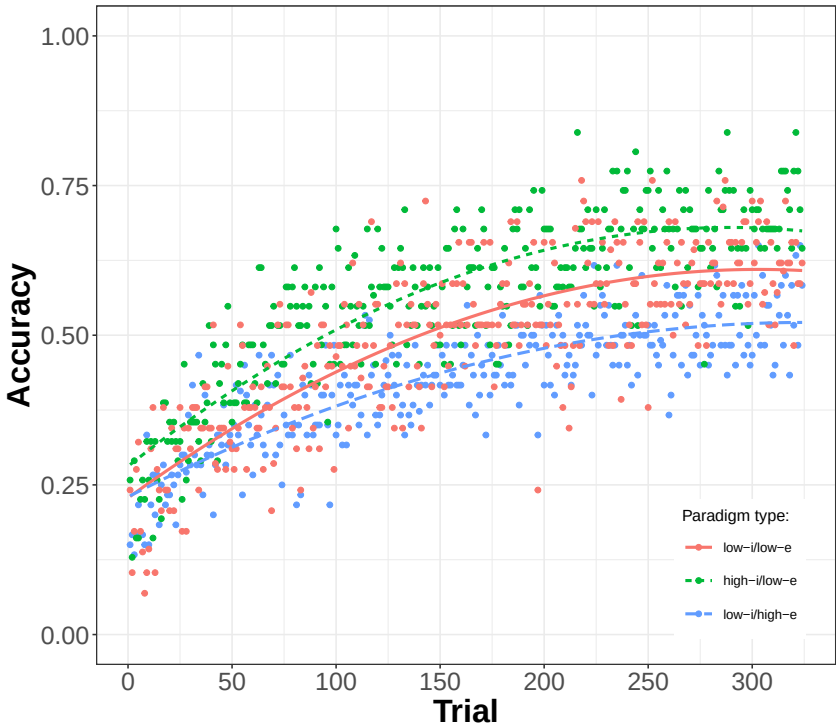


Figure 7: Mean accuracy by trial for each of the three paradigm types (collapsing the two low-i/high-e paradigms). Points indicate the average accuracy across participants for each trial. Lines show quadratic polynomial curves predicting accuracy by trial number for each paradigm type. Learning is worst for the low-i/high-e and best for the high-i/low-e paradigms

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 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 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. Hudson Kam and Newport 2005, 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.

2.3.5 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 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's (2013) 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 and 2.

3.1 *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.

3.2 *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 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 ($r = -0.38$,

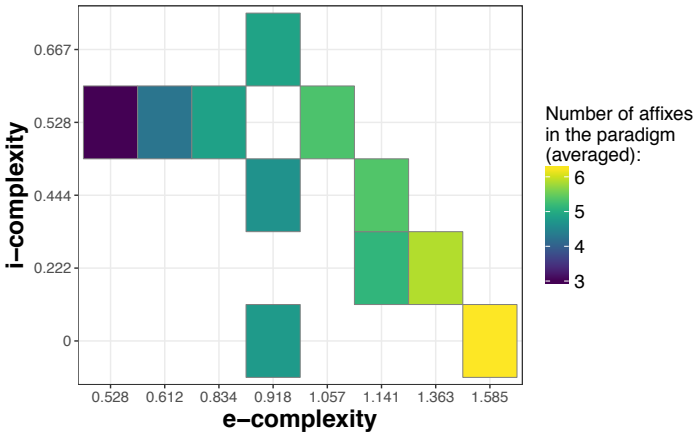


Figure 8: Distribution of randomly generated paradigms in terms of i- and e-complexity. Colour represents the average number of markers for paradigms with specific i- and e-complexity values. No paradigms have high i-complexity *and* high e-complexity. Paradigms with high i-complexity and low e-complexity have on average fewer markers while paradigms with low i-complexity and high e-complexity have more

$t(998) = -13.1$, $p < 0.001$): 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 correlation ($r = -0.94$, $t(421) = -59.24$, $p < 0.001$). Table 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's (2013) findings should be interpreted. They show that across a sample of morphological paradigms in ten languages, e-complexity reaches relatively high 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,[9] and when a paradigm has high e-complexity, it will necessarily also have low i-complexity. Ackerman and Malouf's (2013) findings may therefore at least partly reflect the nature of the relationship between these two

[9] Note however, that the paradigms generated here were matched in size to the paradigms used in Section 2 (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).

Table 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)

(a)

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

(b)

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

measures rather than anything specific to the dynamics of language change.

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

The learning results presented in Section 2 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 2.2.1. 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.

3.3.1 Results

Figure 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.

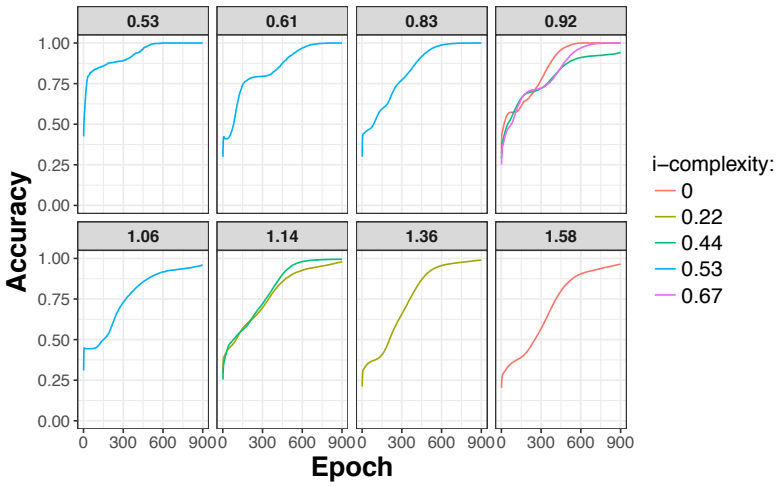
To test how varying values of i-complexity and e-complexity affect learning, we ran a linear mixed-effects regression model predicting summed accuracy by paradigm i-complexity, paradigm e-complexity, the number of different affixes in the paradigm, and their interactions.

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).

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

(a)



(b)

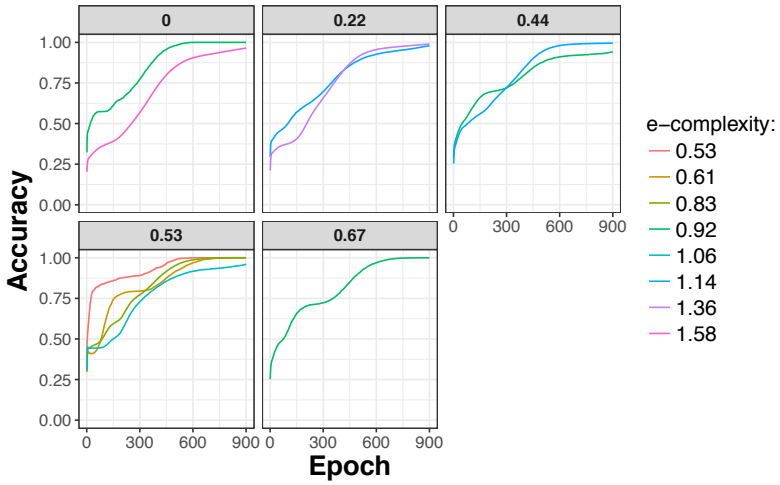


Figure 9: Network learning trajectory for paradigms varying in i-complexity and e-complexity values. (a) i-complexity varying by colour (facet titles showing e-complexity in bits). (b) e-complexity varying by colour (facet titles showing i-complexity in bits). Note that, as discussed above, for some values of e-complexity, the random paradigms do not vary in i-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)

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 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.

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

Table 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

(a)

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

(b)

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

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 ($\beta = -0.034$, $(49992) = -91.4$, $p < 0.001$), 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.

4 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 2 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 *et al.* (2020) and extend them to a more realistic

learning scenario where input includes all forms at all stages (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, compared 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.[10]

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 2, with the two example paradigms differing by types of syncretism, and in Section 3, 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

[10] 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 no specific reason to believe this is the case, one could, in principle, investigate child learners using the kind of study we have reported here.

others (e.g. Pertsova 2012; Maldonado and Culbertson 2019). 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 *et al.* (2020), we tested the effect of i-complexity on generalization with LSTMs, and our results there match Ackerman and 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 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 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, 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. The 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 a larger range of values of the two measures, as opposed to the specific values we tested in Section 2. 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 pervasive-

ness 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. Silvey *et al.* 2015; Kirby *et al.* 2015; Motamedi *et al.* 2019). Since we showed here that e-complexity correlates positively with a num-

ber 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. type/token frequencies of inflection classes or frequencies of forms of grammatical functions in the paradigm).

5 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.

APPENDIX 6

Exploring hyperparameters space 6.1

For the LSTM model presented in Section 2.2 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 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 10–13, and a summary of the mean summed accuracy for all combinations of hyperparameters is presented in Tables 7, 8 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 2.2, namely that in cases where i-complexity has an effect on learning the paradigm, the effect is smaller than the effect of e-complexity.

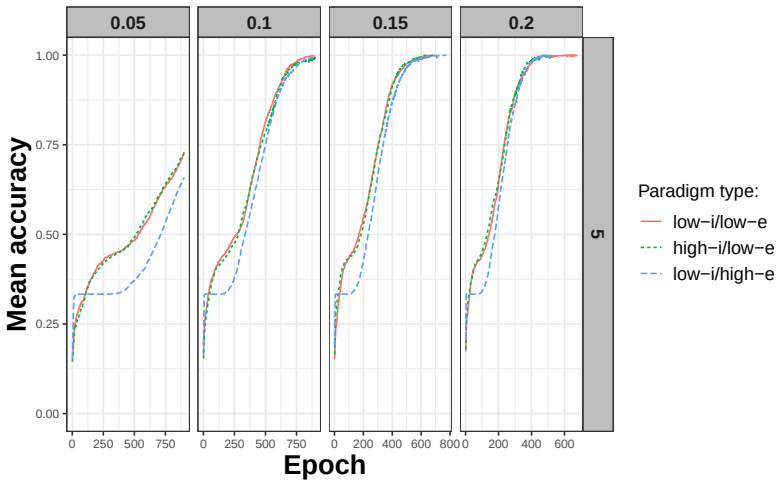
Table 7: Summary of mean of summed accuracy of the model runs optimized with SGD with combinations of hidden and embedding dimensions (5, 25) and learning rates (0.05, 0.1, 0.15, 0.2). Standard deviations are presented in brackets

		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)
	high-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
	/high-e	(41.4)	(51.0)	(33.5)	(21.4)	(41.4)	(26.5)	(18.1)	(13.4)

Table 8: 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

		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)
	high-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
	/high-e	(40.9)	(32.0)	(20.11)	(13.0)	(11.4)	(5.9)	(3.8)	(3.3)

(a)



(b)

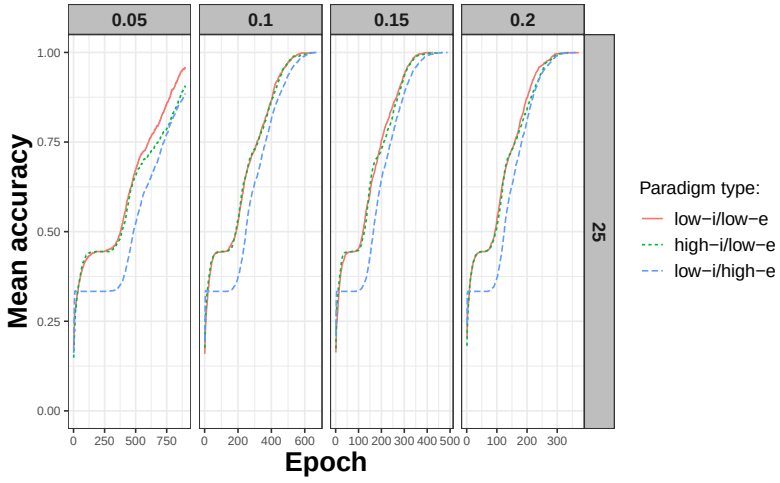
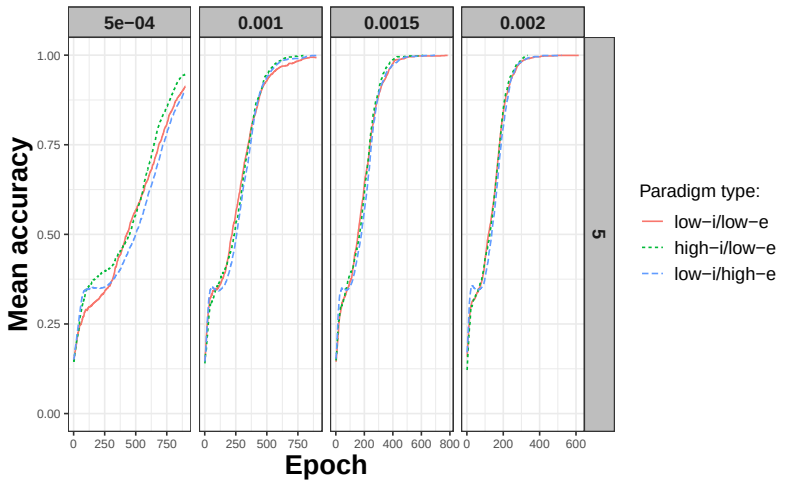


Figure 10: Learning trajectories of networks with two embedding and hidden layer dimensionalities; (a) networks with 5-dimensional embedding vectors and hidden layer, (b) networks with 25-dimentional embedding vectors and hidden layer, trained with different learning rates (columns), and optimized with SGD. *X* axis shows number of epochs up to perfect learning of the forms in the language (differs across learning rates and network dimensions)

(a)



(b)

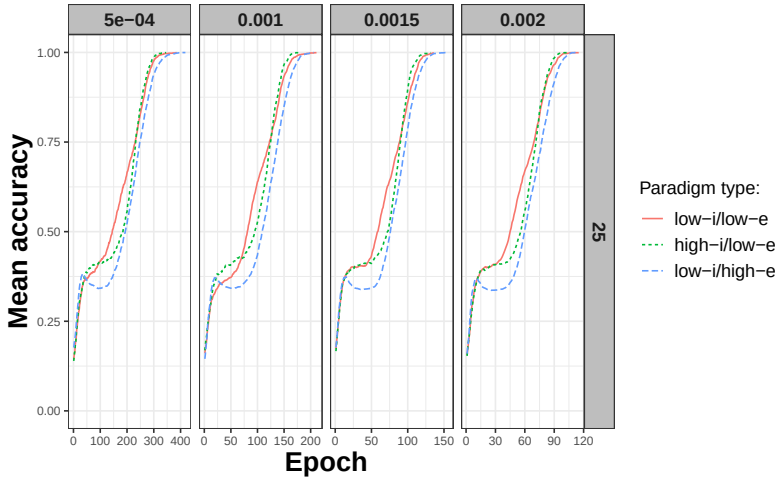


Figure 11: Learning trajectories of networks with two embedding and hidden layer dimensionalities; (a) networks with 5-dimensional embedding vectors and hidden layer, (b) networks with 25-dimentional embedding vectors and hidden layer, trained with different learning rates (columns), and optimized with Adam. *X* axis shows number of epochs up to perfect learning of the forms in the language (differs across learning rates and networks dimensions)

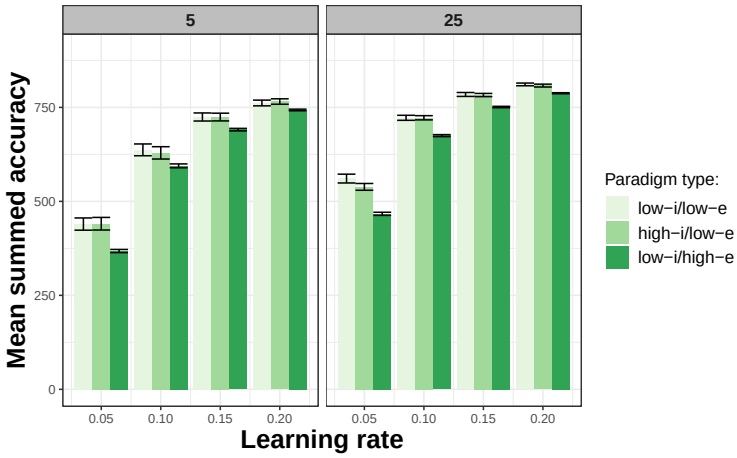


Figure 12: Summed accuracy over the 900 epochs of the networks trained on each of the three paradigm types for models with different learning rates (x axis) and for models with different dimensions (columns) optimized with SGD

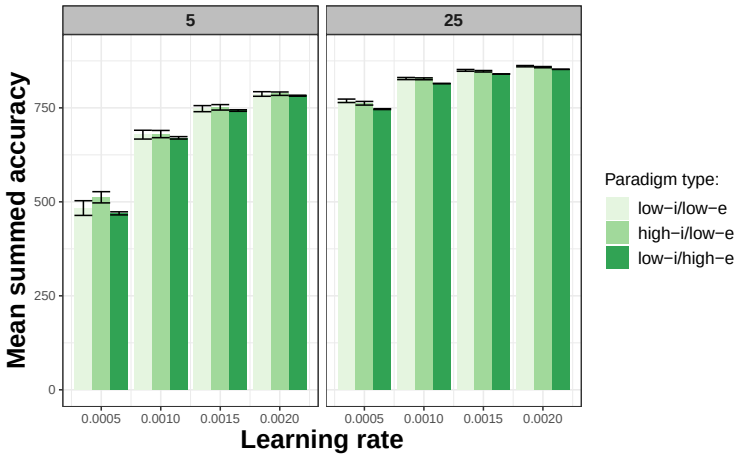


Figure 13: Summed accuracy over the 900 epochs of the networks trained on each of the three paradigm types for models with different learning rates (x axis) and for models with different dimensions (columns) optimized with Adam

REFERENCES

Farrell ACKERMAN, James P. BLEVINS, and Robert MALOUF (2009), Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter, in James P. BLEVINS and Juliette BLEVINS, editors, *Analogy in grammar: Form and acquisition*, pp. 54–82, Oxford University Press, Oxford.

Farrell ACKERMAN and Robert MALOUF (2013), Morphological organization: The low conditional entropy conjecture, *Language*, 89(3):429–464.

Farrell ACKERMAN and Robert MALOUF (2015), The No Blur Principle effects as an emergent property of language systems, in *Proceedings of the annual meeting of the Berkeley Linguistics Society*, volume 41, pp. 1–14.

Ben AMBRIDGE, Evan KIDD, Caroline F. ROWLAND, and Anna L. THEAKSTON (2015), The ubiquity of frequency effects in first language acquisition, *Journal of Child Language*, 42(2):239–273.

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

Matthew BAERMAN, Dunstan BROWN, and Greville G. CORBETT (2005), *The syntax-morphology interface: A study of syncretism*, Cambridge University Press.

Matthew BAERMAN, Dunstan BROWN, and Greville G. CORBETT (2010), Morphological complexity: a typological perspective, `https://www.researchgate.net/publication/266215146`, unpublished manuscript, University of Surrey.

Douglas BATES, Martin MÄCHLER, Ben BOLKER, and Steve WALKER (2014), Fitting linear mixed-effects models using lme4, *arXiv preprint arXiv:1406.5823*.

Balthasar BICKEL and Johanna NICHOLS (2013), Inflectional synthesis of the verb, in Matthew S. DRYER and Martin HASPELMATH, editors, *The world atlas of language structures online*, Max Planck Institute for Evolutionary Anthropology, `https://wals.info/chapter/22`.

James P. BLEVINS (2006), Word-based morphology, *Journal of Linguistics*, 42(3):531–573.

Olivier BONAMI and Sacha BENIAMINE (2016), Joint predictiveness in inflectional paradigms, *Word Structure*, 9(2):156–182.

François CHOLLET *et al.* (2015), keras, `https://keras.io`.

Morten H. CHRISTIANSEN and Nick CHATER (2008), Language as shaped by the brain, *The Behavioral and Brain Sciences*, 31(5):489–509.

Greville G. CORBETT (2009), Suppletion: Typology, markedness, complexity, in Patrick O. STEINKRÜGER and Manfred KRIFKA, editors, *On inflection*, p. 40, Mouton de Gruyter.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner (2019), On the complexity and typology of inflectional morphological systems, *Transactions of the Association for Computational Linguistics*, 7:327–342.

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

Jennifer Culbertson and Elissa L. Newport (2015), Harmonic biases in child learners: In support of language universals, *Cognition*, 139(6):71–82.

Jennifer Culbertson, Paul Smolensky, and Géraldine Legendre (2012), Learning biases predict a word order universal, *Cognition*, 122(3):306–329.

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

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

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

Maryia Fedzechkina, T. Florian Jaeger, and Elissa L. Newport (2012), Language learners restructure their input to facilitate efficient communication, *Proceedings of the National Academy of Sciences*, 109(44):17897–17902.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy (2019), Neural language models as psycholinguistic subjects: Representations of syntactic state, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pp. 32–42.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni (2018), Colorless green recurrent networks dream hierarchically, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pp. 1195–1205.

Sepp Hochreiter and Jürgen Schmidhuber (1997), Long short-term memory, *Neural Computation*, 9(8):1735–1780.

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

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

Tamar Johnson, Jennifer Culbertson, Hugh Rabagliati, and Kenny Smith (2020), Assessing integrative complexity as a predictor of morphological learning using neural networks and artificial language learning,

https://psyarxiv.com/yngw9/, unpublished manuscript, University of Edinburgh.

Michael I. Jordan (1997), Serial order: A parallel distributed processing approach, in John W. Donahoe and Vivian Packard Dorsel, editors, *Neural-network models of cognition*, pp. 471–495, Elsevier.

Charles Kemp and Terry Regier (2012), Kinship categories across languages reflect general communicative principles, *Science*, 336(6084):1049–1054.

Diederik P. Kingma and Jimmy Ba (2014), Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*.

Simon Kirby (2002), Learning, bottlenecks and the evolution of recursive syntax, in Ted Briscoe, editor, *Linguistic evolution through language acquisition*, pp. 173–204, Cambridge University Press.

Simon Kirby, Hannah Cornish, and Kenny Smith (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.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith (2015), Compression and communication in the cultural evolution of linguistic structure, *Cognition*, 141:87–102.

Alexandra Kuznetsova, Per B. Brockhoff, Rune HB Christensen, *et al.* (2017), lmerTest package: tests in linear mixed effects models, *Journal of Statistical Software*, 82(13):1–26.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg (2016), Assessing the ability of LSTMs to learn syntax-sensitive dependencies, *Transactions of the Association for Computational Linguistics*, 4:521–535.

Mora Maldonado and Jennifer Culbertson (2019), Something about ”us”: Learning first person pronoun systems, in *Proceedings of the 41st annual meeting of the Cognitive Science Society*, pp. 749–755.

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

Eric Meinhardt, Rob Malouf, and Farrell Ackerman (2019), Morphology gets more and more complex, unless it doesn't, https://www.researchgate.net/publication/333194657, unpublished manuscript, San Diego State University and University of California San Diego.

Daniel Mirman (2017), *Growth curve analysis and visualization using R*, CRC Press, first edition. edition.

Elliott Moreton and Joe Pater (2012), Structure and substance in artificial-phonology learning, part I: Structure, *Language and Linguistics Compass*, 6(11):686–701.

Yasamin MOTAMEDI, Marieke SCHOUWSTRA, Kenny SMITH, Jennifer CULBERTSON, and Simon KIRBY (2019), Evolving artificial sign languages in the lab: From improvised gesture to systematic sign, *Cognition*, 192:103964–103964.

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

Terry REGIER, Charles KEMP, and Paul KAY (2015), Word meanings across languages support efficient communication, in Brian MACWHINNEY and William O'GRADY, editors, *The handbook of language emergence*, pp. 237–263, John Wiley & Sons, Inc.

Scott SEYFARTH, Farrell ACKERMAN, and Robert MALOUF (2014), Implicative organization facilitates morphological learning, in *Annual meeting of the Berkeley Linguistics Society*, volume 40, pp. 480–494.

Claude Elwood SHANNON (1963), *The mathematical theory of communication*, University of Illinois Press.

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

Catriona SILVEY, Simon KIRBY, and Kenny SMITH (2015), Word meanings evolve to selectively preserve distinctions on salient dimensions, *Cognitive Science*, 39(1):212–226.

Andrea D SIMS and Jeff PARKER (2016), How inflection class systems work: On the informativity of implicative structure, *Word Structure*, 9(2):215–239.

Elizabeth WONNACOTT and Elissa L. NEWPORT (2005), Novelty and regularization: The effect of novel instances on rule formation, in *BUCLD 29: Proceedings of the 29th annual Boston University conference on language development*, pp. 663–673.

Yang XU, Terry REGIER, and Barbara C MALT (2016), Historical semantic chaining and efficient communication: The case of container names, *Cognitive Science*, 40(8):2081–2094.

Noga ZASLAVSKY, Charles KEMP, Naftali TISHBY, and Terry REGIER (2020), Communicative need in colour naming, *Cognitive Neuropsychology*, 37(5-6):312–324.

Tamar Johnson
0000-0003-1071-6750
tamar.johnson@unige.ch

Kexin Gao
kexin.gao@hotmail.com

Kenny Smith
0000-0002-4530-6914
kenny.smith@ed.ac.uk

Jennifer Culbertson
0000-0002-1737-6296
jennifer.culbertson@ed.ac.uk

Centre for Language Evolution,
University of Edinburgh,
Edinburgh, Scotland, United Kingdom

Hugh Rabagliati
0000-0001-9828-5857
hugh.rabagliati@ed.ac.uk

Department of Psychology,
University of Edinburgh,
Edinburgh, Scotland, United Kingdom

Tamar Johnson, Kexin Gao, Kenny Smith, Hugh Rabagliati, and Jennifer Culbertson (2021), *Investigating the effects of i-complexity and e-complexity on the learnability of morphological systems*, Journal of Language Modelling, 9(1):97–150
https://dx.doi.org/10.15398/jlm.v9i1.259