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

Matthew Roberts[1], Luca Onnis[1], Nick Chater[1,2]

[1] Department of Psychology, University of Warwick, Coventry CV4 7AL, UK

[2] Institute for Applied Cognitive Science, University of Warwick, Coventry CV4 7AL, UK

Abstract

The quasi-productivity of natural languages appears to pose two difficult problems for language research. Firstly, why do irregularities in natural language not disappear over time, leaving languages completely regular (a transmission problem), and secondly, how did such irregularity arise in the first place (an emergence problem)? To address the transmission problem, we present an artificial, simplicity-based learner capable of acquiring quasi-regular structures. In doing so, we present an explicitly psychological model of a famously problematic aspect of language acquisition known as Baker's Paradox. We present several simulations of an Iterated Learning Model (ILM) illustrating the emergence and stability of quasi-regular irregularities using a rudimentary language. These simulations offer a possible resolution to the emergence problem. Other possible resolutions are discussed.

Introduction

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

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

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

In such cases we can think of linguistic rules as being quasi-regular: they license the combination of *some* members of syntactic categories, but not others. The difficulty of learning such idiosyncratic absences from partial input and without negative evidence (as is the case with natural language) has become notorious in the language acquisition literature. In particular, given that only a finite set of sentences is ever heard, out of the infinite set of possible sentences in a natural language, it is clear that mere absence of a linguistic form cannot be directly used as evidence that the form is not allowed. Yet, such 'holes' are clearly specific to particular natural languages, and hence cannot be explained by adversion to innate linguistic principles. This problem has been viewed as so severe that it has been labeled Baker's *paradox*; and viewed as raising *logical* problems for the theory of language acquisition (e.g., Baker & McCarthy, 1981)[1].

The approach we adopt here is to apply a general principle of learning to explain how linguistic idiosyncracies can be acquired. Note that the mechanism must be that is sufficiently flexible to capture the huge range of idiosyncrasies across a huge range of linguistic contexts. Moreover, the existence of such a mechanism is required, we contend, to explain the existence of idiosyncracies in language: idiosyncrasies could not have emerged or survived in its absence, as they would have been winnowed out by learning failures by successive linguistic generations. In this respect, Baker's paradox raises a secondary paradox for language evolution. The puzzle of how language acquisition processes can capture what appear to be idiosyncratic 'holes' in the language also raises the puzzle of how difficult-to-acquire linguistic patterns are emerge and are transmitted in the development of languages. Note that, on pain of circularity, whatever learning mechanisms are responsible for learning such idiosyncracies must pre-date the emergence of such idiosyncracies. That is, we cannot view the idiosyncratic nature of language as a stable environment to which biological basis for language acquisition adapted---because without relevant prior learning mechanisms already established, language could not have developed with such idiosyncracies in the first place.

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

The chapter falls into five distinct sections. In the first we discuss the ubiquity of quasi-regular constructions. In the second section we outline why they constitute such an apparently difficult learning problem. We then discuss the relationship between acquisition and evolution, in particular the idea that any hard learning problem of culturally transmitted information entails evolutionary puzzles.

In the third section we present a model that is able to learn quasi-regular structures in a rudimentary language from positive evidence alone, using a very general learning principle: simplicity. The model learns by creating competing hypothetical grammars to fit the language to which it has been exposed, and choosing the simplest. As an explicit metric for simplicity we use Minimum Description Length (MDL), a mathematical idea grounded in Kolmogorov complexity theory (Li & Vitányi, 1997). In acquiring quasi-regular language structures, our model addresses the transmission/acquisition problem.

In the fourth section, we detail several simulations based on an Iterated Learning Model (ILM, e.g. Kirby, 2001) in which a probabilistically generated artificial language is transmitted over 1,000 generations of simplicity-based learners. The results of these simulations chart not only the stability but also the emergence of quasi-productivities in the language. In particular we show that:

a) Exceptions are stable across successive generations of simplicity driven learners.
b) Under certain conditions, statistical learning using simplicity can account for the emergence of quasi-productivity in a language.

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

Baker's Paradox and linguistic quasi-productivity.

A mainstay of linguistic analysis has been that human languages are composed of a limited number of basic units (features, segments, syllables, morphemes, words, phrases, clauses, etc.) that can be combined by a small number of generative rules to create larger units. Postulating the existence of recursive rules allows for an infinite number of sentences to be created. This generativity goes well beyond theoretical linguistic description, as it is typically taken to be embodied in the psychological mechanisms responsible for acquiring and representing linguistic rules and units.

Although the capacity to generalize from a limited set of examples to novel instances is an uncontroversial aspect of the human cognition, a puzzle that has attracted linguists is that natural languages, although productive, are never fully regular. There appear to be finely-tuned lexical and syntactic selectional

constraints that native speakers are aware of. Expected regular structures may either be replaced (e.g. *went* for **goed*) or they may be disallowed completely. These semi-productive structures may be seen as a special case of irregularity where the irregular form is absent, i.e. there seems to be an unfilled slot that constrains open-ended productivity. Consider, for instance, a transformational rule such as *to be* Deletion (after Baker, 1979):

(1) *X – to be – Y*

→ *X, 0, Y*

(2) *The baby seems/appears to be happy*

(3) *The baby seems/appears happy*

(4) *The baby seems to be sleeping*

(5) *The baby happens to be sleepy*

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

(7) **The baby happens sleepy*

On the basis of positive evidence positing the transformational rule in (1) is misleading with regard to the perfectly plausible but ungrammatical predictions that it gives about sentences (6) and (7). Such 'unfilled slots' cannot be accounted for by the general rule. Similarly, consider the lexical constraints on the collocations between, for instance, adjective and noun below:

(8) *strong/ high/*stiff winds*

(9) *strong /*high/*stiff currents*

(10) *strong/*high/stiff breeze*

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

(11) $NP_1 – V – NP_2 –$ *to* NP_3

→ NP_1, *V*, NP_3, NP_2 (optional)

(12) *We sent the book to George*

(13) *We sent George the book*

(14) *We reported the accident to the police*

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

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

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

The logical problem of language acquisition

At a general level, the so-called logical problem of language acquisition is that learning a language from experience alone is impossible because linguistic experience is too incomplete and contradictory. In the first place, a learner observes only a limited set of the infinite number of utterances in his/her language. From this, he/she must distinguish a certain set of 'grammatical' utterances among all the other utterances that he has never heard and may never produce. The problem is particularly acute when considering the case of quasi-productivities, which yield Baker's Paradox (also known as the Projection problem) after Baker (1979). Baker noted that quasi-productive regularities such as those above pose a genuine puzzle for any account of language acquisition. This is principally because the unfilled slots they create in the language occur *within* the space of allowable sentences and nonetheless are somehow blocked by language learners. A crucial tenet of the logical problem is that indirect negative evidence in the form of absence is not sufficient to constrain the learner's hypotheses about the correct grammar, because there are many linguistic sentences that a learner has never heard but are nonetheless grammatical (Pinker, 1994). There are therefore many hypothesis grammars that

would be consistent with the positive data available. It is suggested that such a hard learning problem necessitates the existence of powerful innate linguistic tools. However, the paradox raised by Baker is that even postulating a Universal Grammar that restricts the search space for potential grammars does not solve this particular problem, since unfilled slots are highly idiosyncratic across languages. We contend that because these constructions cannot be derived from universal principles, they must be determined by the learner on the basis of exposure to the language, thus providing a solution to Baker's paradox. Nor is it the case that this apparently intractable computational problem will disappear in the face of simple appeal to semantics. For instance, it is often claimed that transitive and intransitive verbs may be distinguished by virtue of the fact that transitive verbs refer to sequences involving both agents and patients, whilst intransitives involve only agents:

(16) *John broke the cup*

(17) *The cup broke*

(18) *John kissed Mary*

(19) **Mary kissed*

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

(20) *John waved Mary goodbye*

(21) *John waved goodbye to Mary*

(22) **John said Mary hallo*

(23) *John said hallo to Mary*

or, again from Baker:

(24) *It is likely that John will come*

(25) *It is possible that John will come*

(26) *John is likely to come*

(27) **John is possible to come*

Hence, we argue that some degree of arbitrariness must be accounted for in quasi-regular constructions (see also Culicover, 1999, on the case for at least partial independence of syntax from semantics in the case of unfilled slots). If idiosyncrasy is to be found at the core of grammar and can neither be accounted for by universal principles nor semantically determined completely, it must be learnable from experience. Before we consider how such learning might occur, we consider why the existence of this acquisition problem entails two equally puzzling evolution problems.

The logical problem of language evolution

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

A solution to the acquisition problem

In the following section we first outline the simplicity principle and Minimum Description Length (MDL) as a metric for simplicity. We then describe how this forms the basis of a language learning mechanism, detail a rudimentary toy language for use in simulations and give details of a single learner simulation in which a learner agent learns idiosyncratic exceptions. Using this machinery, we show, for this rudimentary language, how it is possible to acquire quasi-regular patterns in a language, from positive evidence, by preferring the grammar that corresponds to the simplest representation of the corpus that has been encountered. Roughly, the simplicity principle allows learners to determine when absence of a particular construction from the corpus can be taken as genuine evidence that it is absent disallowed. This general type of approach to Baker's paradox has been discussed by a range of authors (Dowman, 2000; Stolcke,1994).

The Simplicity Principle and MDL

From an abstract point of view, learning from experience can be thought of as finding patterns in a finite set of data. Any finite set of data is consistent with an infinite number of possible patterns; the problem is how to choose between them. The simplicity principle (e.g. Chater, 1999) asserts that the cognitive system will always prefer simpler patterns over more complex ones. The mathematical theory of Kolmogorov complexity provides an elegant way to think about simplicity through two key insights (Li & Vitányi, 1997): Firstly, that the length of the shortest program in a universal programming language (e.g., any conventional computer language, such as C++ or Java) that regenerates that object is a natural measure of the complexity of that object. Secondly, the length of that program is independent of the choice of the specific universal programming language, up to a constant. This length is known as the Kolmogorov complexity of an object. This approach has yielded a rich mathematical literature (Li & Vitányi, 1997), and a number of theoretical results concerning language learnability from positive evidence (e.g. Chater and Vitányi , 2001), but is problematic from a practical point of view, in that Kolmogorov complexity itself is uncomputable. This provides the motivation for practical variants of this approach, where complexity is measured using restricted statistical coding schemes, rather than the full power of a universal programming language (Rissanen, 1987, 1989; Wallace & Freeman, 1987). Note that it is not necessary to discover the *actual* binary encoding of an object: it is possible to deal only in code *lengths*, i.e. the number of bits necessary to describe an object or event. This figure can be specified if a probability can be associated with that object of event (Shannon, 1948), using standard information theory. Highly probable or frequent objects or events are associated with short (simple) encodings.

Simplicity-Based Language Learning: The Learner as Gambler

The simplicity principle, outlined above, demonstrates how the simplest model of experience can be thought of as that represented by the shortest binary code. In this instance the binary code must represent two things: firstly a hypothesis, or grammar, that describes the language to which the learner is exposed. Secondly, all the language that has been heard must be represented *under the hypothesis*. This may be expressed formally:

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

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

hypothesis. The length of code necessary to represent data will differ between hypotheses.

Our model of the learner does not aquire vocabulary or induce categories and rules from scratch. We take productive rules to be already learned. Thus our model is already at the stage at which children make over-general errors. The task is to spot which of the constructions allowable under the rule are in fact blocked---to find the holes in the language. Learning proceeds by a series of "gambles." The learner bets that a particular construction is not allowed and that it will therefore never be encountered. In making this gamble it must specify the construction as part of a new hypothesis, H. Coding this specification requires some bits of information, so the complexity of the new hypothesis increases. However, the learner has reduced the number of allowable constructions that it can expect to encounter. It has therefore increased the probability of those remaining. The number of bits required to specify future data under the new hypothesis is therefore reduced. Thus, if it is true that the construction is not allowed, the learner will gradually win back the number of bits that it gambled in specifying the exception. As more language is heard the new hypothesis will eventually come to be associated with a shorter codelength than the original. If the gamble is inappropriate, however, the learner will encounter a construction that it has wrongly presumed to be disallowed. This is associated with a probability of 0, and hence an infinite code-length, so the 'gamble' is abandoned. Our model generates a new hypothesis every time it gambles on a particular construction, with all hypotheses running in parallel. The preferred hypothesis is always that associated with the shortest codelength.

Onnis, Roberts & Chater (2002) show that a batch learner (i.e., a learner that runs all calculations, after the entire corpus has been encountered) employing this strategy is able to distinguish genuine constructions from blocked ones as a result of exposure to data from the CHILDES database of child directed speech (MacWhinney, 2000). Here, we implement an online version that was able to postulate exceptions and create new hypotheses during the course of exposure to a rudimentary toy language. Algorithmic details are given in Appendix A; the following two sections describe the toy language and the learner's ability to discover exceptions in it.

Learning a rudimentary language

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

S_1->AB,

S_2->BA,

A->{a_1,a_2,a_3,a_4},

B->{b_1,$b2$,b_3,b_4},

*->{(a_2),(a_2b_2),(a_2b_3),(a_2b_4), (b_1a_1),(b_2a_1),(b_3a_1),(b_4a_1)} [3]

where the examples generated by * are blocked. This language can mimic the pattern of alternations, for example transitive and intransitive verb constructions. In English, verbs can nominally occur in either a transitive or an intransitive context, but some are blocked from occurring in one or the other. This is analogous to the patterns in our toy language, where items in either category may in principle occur in either the first position, but can be blocked from doing so by entries in the exceptions element. This is illustrated in Figure 1.

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

AB	***BA***
a_1B	*$*Ba_1$*
$*a_2B$	Ba_2
a_3B, a_4B	Ba_3,Ba_4

Figure 1. The structure of the toy language mimics that of Baker's Paradox for alternations. a_1 and a_2 could be blocked from occurring in BA and AB constructions respectively by entries in the exceptions element such as a_2b_1, a_2b_2 or b_1a_1, b_2a_1 etc. For the first generation agent in each simulation, however, all As occurred in both contexts (that is, they were 'alternating'). 'Cut', 'fall', and 'break' are examples of alternating and non-alternating verbs. Levin (1993) provides an extensive list of alternations in English.

Samples of the language were produced by a parent agent and experienced by a learner agent. We assume that parents and learners share knowledge of word frequency. This allows both to associate each word with a probability of occurrence. Sentence probabilities are taken to be the product of two probabilities: that of the first word and that of the second word, given the first. Parent agents use these probabilities to produce samples of the language stochastically. Learners use them to calculate codelengths (in bits) for different hypotheses. We assume that word frequencies are distributed according to Zipf's law (Zipf, 1948), an ubiquitous power law distribution in natural language (Bell, Cleary & Witten, 1990): If we rank words in terms of frequency, then frequency of any word is the inverse of its rank. Details are given in Appendix A.

Learner agents begin with a single, completely regular hypothesis about the language i.e., all sentences are allowed. This is equivalent to [3] with the exceptions element empty. As they experience samples of the language, the learner agents compare the probability of each sentence with the total number of sentences they have heard. A new hypothesis is generated if the total exceeds a threshold (where the threshold is a

function of sentence probability; thus the threshold is different for each sentence). Each new hypothesis is simply a clone of the most recent hypothesis to be generated (or the original, if it is the first) with the addition of the sentence in question to the exceptions element. This addition entails an increment in the codelength associated with the new hypothesis, and a re-scaling of the probabilities for the remaining sentences.

Each sentence encountered entails an increment in the number of bits associated with each hypothesis, but since the creation of a new hypothesis involves rescaling sentence probabilities, this increment differs between hypotheses. All algorithmic details are given in Appendix A. Figure 2 illustrates the codelengths associated with all the hypotheses entertained by a learner agent after exposure to 50 sentences of a language containing 11 exceptions.

[insert figure 2 about here]

Figure 2. The codelength (number of bits) associated with each hypothesis grammar entertained by a learner after exposure to 50 sentences of a language containing 11 exceptions. The shortest codelength is obtained by the 12^{th} hypothesis (the first contains no exceptions).

Figure 2 illustrates that the learner agent creates many hypotheses, but that the shortest codelength is associated with the one that matches the language to which it was exposed. It is important to note that the sentence comprised of the two least frequent words was associated with a probability of approximately 1/100. It was therefore highly unlikely that it would have occurred in a corpus of 50 sentences. In addition, the learner received no feedback on its learning, other than more samples of the language. These conditions mirror to a modest extent the 'poverty of the stimulus', according to which children never hear all the possible sentences of a language and do not typically receive explicit negative feedback. In addition, our language contains, of course, no semantics and has no communicative function: we do not attempt to model the relationship between meanings-signals-referents nor try to give functional explanations of language change as in other models. In

general, however, part of the fascination of the constructions investigated here is that their idiosyncrasy does not seem to be primarily semantically or functionally determined.

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

	***a*$_1$**	***a*$_2$**	***a*$_3$**	***a*$_4$**	***b*$_1$**	***b*$_2$**	***b*$_3$**	***b*$_4$**
a_1	*	*	*	*				
a_2	*	*	*	*	*	*	*	*
a_3	*	*	*	*				
a_4	*	*	*	*				
b_1	*				*	*	*	*
b_2	*				*	*	*	*
b_3	*				*	*	*	*
b_4	*				*	*	*	*

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

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

Language Learning over Generations - ILM simulations

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

agent per generation, so issues of population dynamics are neglected. All agents were "genetically" homogeneous, i.e. all were equipped with identical learning facility and started from the same point in their development. The first generation agent was exposed to probabilistically generated samples of the completely regular toy language used in the single-learner simulation. Subsequent agents were exposed to probabilistically generated samples of the language as learned by the preceding generation. Although complete regularity at the outset is probably unrealistic, our intent is not so much to replicate an historic development of languages as to test the conditions for the emergence and stability of irregularities. We test this in the least favourable condition for their emergence, i.e. an ideal fully regular language.

The mean number of sentences heard by each agent was the same within each simulation, but varied between simulations. In different simulations, successive generations of learners heard between 25 and 65 sentences. Again, it was unlikely that any agent was exposed to all the sentences in the language, and agents received no negative feedback on their learning. When an agent had been exposed to the required number of sentences, one hypothesis entertained by that agent was selected. This hypothesis was then used as the basis for generating the sentences that would be heard by the succeeding generation. The hypothesis chosen was always that associated with the simplest interpretation, i.e., that with the shortest code length.

Results

Figure 4 charts the emergence and stability of exceptions in four simulations. The number of sentences heard by each generation was critical to both. Where each generation heard a short corpus (mean number of sentences, n, of 30, Fig. 4(a)), exceptions frequently emerged but were highly unstable: they rarely remained in the language for more than a few generations. With a long corpus (mean n=60, Fig. 4(d)) exceptions were less likely to emerge; in contrast to Figs 4(a) - 4(c) no exceptions emerged for almost 400 generations. However, once they had emerged they were much less likely to be lost from the language than with shorter corpora.

[insert Figure 4 about here]

Figure 4. The number of exceptions contained in the language transmitted by each learner to the

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

Figure 4 suggests that exceptions are posited during the early stages of language acquisition. With a relatively small amount of data, learners may postulate that the language contains many exceptions that do not in fact exist. As more data becomes available, such early hypotheses are either confirmed or exposed as spurious. These simulations suggest a trade off between emergence and stability of exceptions. The crucial factor mediating this trade off is the amount of language heard by each generation. If each generation hears a great deal of data, exceptions are unlikely to emerge: any that are posited will later be shown to be false. However, if exceptions are to be stable, each generation must hear enough language to learn the exceptions that existed in the previous generation.

Discussion and conclusion

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

**I gave my mummy it*

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

The quest for simplicity is hardly a new idea and appears, for instance, in the early works of Chomsky (1955; 1965: 25): under the notion of markedness the grammar being constructed directly reflects the linguistic input. If the input contains information that points to a certain complex grammatical relation, the learner will acquire it, but if the input lacks such information, the principles that govern generalization will prevent the learner from constructing the more complex grammar. The markedness approach was abandoned in generative linguistics, in part because of the lack of a metrics for establishing the simplicity of grammars, and partially for the rise of the 'poverty of the stimulus argument' whereby linguistic experience seems hopelessly unreliable. Such caveats are dealt with in this paper: firstly, the MDL approach provides a quantitative metric for simplicity; secondly, the poverty of the stimulus instantiated in the transmission bottleneck seems a necessary precondition for the emergence of exceptions rather than a hindrance to language evolution. There is a critical size for the bottleneck: too little or too much exposure to the language fails to yield stable patterns of quasi-productivity[3].

Another defining feature of the simulations described in this chapter is that they rely on word frequencies to assign probabilities to sentences. We have also assumed that the distribution of word frequencies follows Zipf's law (Zipf, 1948). These assumptions merit some discussion. There are two important reasons for applying a power law distribution to word frequency: firstly, it has been shown in the past to have important implications for the emergence of irregularities in ILM simulations of language evolution (e.g. Kirby, 2001), and secondly, such frequency distributions are ubiquitous in natural language.

Kirby (2001) has shown that benign irregularities[4] will spontaneously emerge in compositional language structure if frequency distributions follow Zipf's law. When this is the case, the very frequent phrases at the 'head' of the distribution are shortened to irregular forms, resulting in selection under a similar MDL metric as that described here. This phenomenon does not appear to occur when frequencies do not follow a power law. We can see the impact of Zipf's law on our simulations by considering the likely results if word frequencies had been evenly distributed (i.e. if all sentences had been assigned equal probability). In such a case, the threshold number of sentences for learning a particular exception would have been the same for every sentence. Thus the learner would either encounter enough sentences to learn all the exceptions at once, or would not learn any exceptions at all. Any sentences not encountered before the threshold was reached would be posited as exceptions. It is not impossible that exceptions would emerge and survive under such conditions, but it seems unlikely that we would see the patterns of emergence and stability outlined above.

In following Zipf's law, the frequency distribution of words in our toy language mirrors that found in natural languages: word frequencies in natural language text and corpora follow such distributions quite precisely, as do a number of other natural language statistics (Bell *et al.*, 1990). The assumption that the probability of a given sentence is perceived as a function of word frequencies is more controversial. It seems highly unlikely that this would be exclusively the case in natural language; we would be surprised if factors such as semantics and phonology did not play a role. However, no factors other than the frequency and collation statistics were available in our language. We contend that it is a plausible assumption that these factors also play a role in determining our perceptions of the probability of a particular sentence occuring. We speculate that in the absence of other factors they must determine them exclusively.

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

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

In this chapter we have shown that a potentially hard problem of language acquisition, that of quasi-regularity, gives rise to a paradox of language evolution. We have shown that the acquisition problem may be solved by incorporating a learning bias towards simplicity. This solution goes some way towards resolving a

related paradox in language evolution: given sufficient exposure to samples of language, quasi-regular structures are learnable, and hence stable over generations. In addition to this we have shown that under some conditions, quasi-regular structures may emerge in a language even if it were initially completely productive. However, we make no assumptions as to the origins of language in the human species. The starting point of a fully regular language should not be taken as an hypothesis about historical languages. It rather served the purpose of demonstrating that quasi-regular structures may emerge spontaneously, and hence constitute a natural stable equilibrium for languages across time.

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

References

Baker, C.L. (1979) Syntactic theory and the projection problem. *Linguistic Inquiry 10*: 522-581

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

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

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

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

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

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

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

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

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

Chomsky, N. (1957). *Syntactic structures*. Mouton: The Hague.

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

Culicover, P. W. (1999). *Syntactic nuts*. Oxford: Oxford University Press.

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

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

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

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

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

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

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

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

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

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

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

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. ??

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

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

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

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

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

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

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

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Prasada, S., & Pinker, S. (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes*. 8: 1-56

Quinlan, J. R. & Rivest, R. (1989). Inferring decision trees using the minimum description length principle.

Information and Computation, 80, 227-248.

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

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

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

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

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

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

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

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

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

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

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

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

APPENDIX A

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

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

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

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

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

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

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

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

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

Exceptions: b_1,a_2 b_2,a_2 b_3,a_2 b_4,a_2

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

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

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

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

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

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

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

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

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

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

In these simulations not all sentences were equally probable. Less probable (and absent) sentences should require more language to be encountered before they could be considered exceptions. This was taken into account by making use of a general derivation (not specific to these simulations) based on the premise that an exception should be postulated at the point at which the investment of bits necessary to specify it would have been recouped had it been postulated before any language was heard.

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

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

If we assume that the probability of any particular sentence is small (i.e. near 0), a Taylor expansion gives that

$\log_2\left(\frac{1}{1-p_{(x)}}\right)$ approximately equals $p_{(x)}$. From this we can conclude [7].

---

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

2 our thanks to an anonymous reviewer for providing this example.

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

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