Simplicity: A Cure for Overgeneralizations in Language Acquisition?

Luca Onnis (l.onnis@ warwick.ac.uk)
Department of Psychology, University of Warwick
CV4 7AL Coventry, UK

Matthew Roberts (m.roberts.2@ warwick.ac.uk)
Department of Psychology, University of Warwick
CV4 7AL Coventry, UK

Nick Chater (nick.chater@warwick.ac.uk)
Department of Psychology and Institute for Applied Cognitive Science, University of Warwick
CV4 7AL Coventry, UK

Abstract

A formal model of learning as induction, the simplicity principle (e.g. Chater & Vitányi, 2001) states that the cognitive system seeks the hypothesis that provides the briefest representation of the available data– here the linguistic input to the child. Data gathered from the CHILDES database were used as an approximation of positive input the child receives from adults. We considered linguistic structures that would yield overgeneralization, according to Baker's paradox (Baker, 1979). A simplicity based simulation was run incorporating two different hypotheses about the grammar: (1) The child assumes that there are no exceptions to the grammar. This hypothesis leads to overgeneralization. (2) The child assumes that some constructions are not allowed. For small corpora of data, the first hypothesis produced a simpler representation. However, for larger corpora, the second hypothesis was preferred as it lead to a shorter input description and eliminated overgeneralization.

Introduction

Overgeneralizations are a common feature of language development. In learning the English past tense, children typically overgeneralize the '-ed' rule, producing constructions such as *we holded the baby rabbits* (Pinker, 1995). Language learners recover from these errors, in spite of the lack of negative evidence and the infinity of allowable constructions that remain unheard; it has been argued that this favours the existence of a specific language-learning device (e.g. Chomsky, 1980; Pinker, 1989). This is an aspect of the 'Poverty of the Stimulus' argument. We report on a statistical model of language acquisition, which suggests that recovery from overgeneralizations may proceed from positive evidence alone. Specifically, we show that adult linguistic competence in quasi-regular structures may stem from an interaction between a general cognitive principle, *simplicity* (Chater, 1996) and statistical properties of the input.

According to Baker's Paradox (Baker, 1979) children are exposed to linguistic structures that they subsequently overgeneralize, demonstrating that they capture some general structure of the language. However, some generalizations are grammatically incorrect and children do not receive direct negative evidence from caretakers (e.g. corrections labeling such overgeneralizations as disallowed). The paradox is that non-occurrence is not in itself evidence for the incorrectness of a construction because an infinite number of unheard sentences are still correct. The irregularities that Baker referred to can be broadly labeled *alternations* (Levin, 1993; see also Culicover, 2000). For instance, the dative alternation in English allows a class of verbs to take both the double-object construction (*He gave Mark the book*) and the prepositional construction (*He gave the book to Mark*). Hence the verb *give* alternates between two constructions. However, certain verbs seem to be constrained to one possible construction only (*He donated the book to Mark* is allowed, whereas **He donated Mark the book* is not). Such verbs are non-alternating. From empirical studies we know that children do make overgeneralization errors that involve alternations, such as **I said her no* (by analogy to *I told her no*, Bowerman, 1996; Lord 1979).

In this paper we present alternation phenomena from the CHILDES database (MacWhinney, 2000) of child-directed speech which will be used in the computer model. Secondly, we introduce the *simplicity principle* (Chater, 1996), based on the mathematical theory of Kolmogorov Complexity (Kolmogorov, 1965). Thirdly, we present an artificial language designed to model the CHILDES data, and describe simplicity-based models of language processing and the simulations of recovery from overgeneralizations. Lastly we discuss the limitations of this specific model and some implications for research on language acquisition.

Causative alternations in child-directed speech

Suppose we have a language in which verbs belong to three distinct classes (V1, V2, V3). Each class is related to two syntactic contexts (C1, C2). One class of verbs (V1) appears in both contexts. Two other classes of verbs (V2 and V3) occur in one context only. We can produce a simple table to visualize the alternation:

Table 1: Alternating and non-alternating verbs across contexts

	C1	C2
V1	1	1
V2	0	1
V3	1	0

The causative alternation in English is of this kind. Verbs like *break* behave both transitively (*I broke the vase*) and intransitively (*The vase broke*), whereas verbs like *disappear* behave only intransitively (*The rabbit disappeared* is allowed; but **I disappeared the rabbit* is not) and verbs like *cut* are found only in transitive contexts (**The bread cuts* is not allowed). An analysis of CHILDES revealed that verbs in child-directed speech fit the pattern of the above idealization: a number of verbs are exclusively transitive or intransitive (see Table 2).

Children eventually generalize the structures of the language they are exposed to. A typical generalization occurs when children say *Don't you fall me down* (Bowerman, 1982; Lord, 1979). This is an overgeneralized use of a non-causative verb as causative. In the causative construction, some verbs like *break* can be used both transitively with a semantic element of cause (*I broke the vase*) and intransitively (*the vase broke*). Verbs like *break* alternate between two constructions. However, *fall* can only be used intransitively, and *hear* only transitively. The acquisition of verbs' argument structure seems particularly complicated as the way verbs behave syntactically is largely arbitrary. Semantically similar verbs like *say* and *tell*, or *give* and *donate* allow for different constructions.

Bowerman (1982) and Lord (1979) recorded a total of 100 different cases in which two-argument verbs are used with three arguments (e.g. *You can drink me the milk*). The developmental literature suggests that when children acquire a new verb they use it productively in both constructions, without specific directional bias (Lord, 1979). It is also worth noting that alternations can be theoretically distinguished from other forms of

Table 2: Verbs in child-directed speech occurring in transitive and intransitive contexts pooled from the CHILDES English sub-corpora (MacWhinney, 2000).

	Verb	Transitive occurrences	Intransitive occurrences
Category V1	bounce	75	117
	break	1251	268
	burn	86	60
	close	855	56
	freeze	18	61
	grow	59	330
	move	966	560
	open	1590	232
	pop	104	153
	rip	139	9
	roll	405	164
	shake	147	26
	slide	65	120
	swing	38	96
	tear	167	20
	turn	2690	600
Category V2	arrive	0	41
	come	0	18437
	dance	0	370
	die	0	141
	disappear	0	73
	fall	0	2945
	go	0	65193
	rise	0	14
	run	0	1569
	stay	0	1413
Category V3	bring	3028	0
	cut	1315	0
	drop	640	0
	kill	120	0
	lift	392	0
	push	1609	0
	put	27154	0
	raise	25	0
	take	9724	0
	throw	2090	0

irregularization like the irregular past tense. In the case of *goed-went* for example, recovery from the overgeneralized form **goed* can be accounted for by directly invoking a competition strategy (MacWhinney, 1987): as the number of *went* in the input increases, it will win over the irregularized form *goed*, which has 0 frequency in the input. Alternations are interesting theoretically in that the competition model does not seem applicable for these. The overgeneralized form does not have an irregular alternative: there is simply a "hole" in the language. This argument was raised by Baker in his distinction between benign exceptions (like

the past tense) and truly problematic alternations like the ones we consider here (Baker 1979).

For the purpose of showing how such problematic irregularities can be learnt using a simplicity principle, we take the causative alternation described above as a working example. We extracted verb frequencies from the CHILDES Database. CHILDES contains a total of nearly ten million words of child-directed speech. Because we are interested in showing that the input the child receives is rich enough for recovery of overgeneralization by induction, only the adult speech in the corpus was selected and analysed.

Simplicity and Language

The simplicity principle (Chater, 1996) states that in choosing among potential models of finite data, there is a general tendency to seek simpler models over complex ones and optimize the trade-off between model complexity and accuracy of model's description (i.e. fit) to the training data. Complexity is thus defined as:

$$C = C(\text{model}) + C(\text{data}|\text{model})$$

The favoured model of any finite set of data will be that which minimizes this term.

In order to compare different grammars we need a measure of simplicity and a "common currency" for measuring both the model complexity and the error term complexity. Fortunately this is possible by viewing grammar induction as a means of *encoding* the linguistic input; the grammatical organization chosen (the "knowledge" of the language) is that which allows the simplest encoding of the input. A tradition within mathematics and computer science, Kolmogorov complexity, shows that the simplest encoding of an object can be identified with the shortest program that regenerates the object (Li & Vitanyi, 1997).

Every sentence generated from a lexicon of *n* words may be coded into a binary sequence. The length of a message refers to a binary string description of the message in an arbitrary universal programming language. The binary string can be seen as a series of binary decisions needed to specify the message; smaller lengths correspond to simpler messages. The brevity of an input A_i is associated to its probability $P(A_i)$ of occurrence. Shannon's (1948) noiseless coding theorem specifies that:

$$\text{Length} = \text{Log}_2[1/P(A_i)]$$

More probable events are therefore given shorter codes. Li & Vitanyi (1997) have shown that the length K(x) of the shortest program generating an object x is also related to its probability Q(x) by the following *coding theorem*:

$$K(x) = \log_2[1/Q(x)]$$

Finally, the *invariance theorem* (Li & Vitanyi, 1997) assures that the shortest description of any object is *invariant* (up to a constant) between different universal languages, thus granting a measure of simplicity that is independent of the data and of the programming language used to encode the data. The above formalizations allow us to replace "Complexity" with "Length" and state that "the best theory to infer from a set of data is the one which minimizes the length of the theory and the length of the data when encoded using the theory as a predictor for the data" (Quinlan and Rivest, 1989; Rissanen, 1989). It is important to note that whilst the MDL principle is well established as a machine learning tool for grammar induction, such models typically make use of parsed corpora or other psychologically implausible inputs (e.g. Osborne, 1999). This paper uses MDL as a metric to present simplicity as a specifically psychological principle.

Modelling language learning with simplicity

In any study of grammar induction, and in particular in the simplicity framework, it is crucial to see a grammar as a *hypothesis about the data*. The best hypothesis is the one that compresses the data maximally, so we can also think of a grammar as compression of the data. We can see the achievement of adult linguistic competence as a process of building different hypotheses about the language in order to achieve optimum compression. The essence of compression is to provide a shorter encoding of the data, enabling generalizations and correct predictions. Alternations are particularly informative about the possibility of a cognitive system to capture dependencies from limited data. If linguistic structures were completely regular, then generalizing from a few data would be easy. But as alternations are quasi-regular, meaning there are exceptions to their regularity, a learner must capture fine dependencies in order to generalize whilst avoiding overgeneralizations.

The issue is to choose the candidate model of the right complexity to describe the corpus data, as stated by the simplicity principle. We can compare different hypotheses (grammars) at different stages of learning and choose, for each stage, the one that minimizes the sum of the grammar-encoding-length and the data-encoding-length. In the following section we compare data compression of corpora by two similar models. The difference between them is that one posits a completely regular rule, whilst the other posits a regular rule and some exceptions to it. We can think of the second model as having 'invested' in exceptions. Each exception initially produces less compression overall,

since the exceptions cost some bits to specify. However, each exception shortens the code-length for each item in the corpus, and the second model thereby 'recoups' its investment over time.

The Models

This approach to language acquisition does not focus on how learning occurs. Rather, these simulations run several models concurrently to show that the rate of increase of code-length differs between hypotheses about language. This section describes the structure of two hypotheses (grammars); the first gives rise to overgeneralization phenomena whilst the second does not. These were designed in conjunction with a very simple artificial language, which was subsequently used to test the models. A brief outline of the language is given here to facilitate the description of the model. A more detailed consideration of how the artificial language relates to data from corpora of child-directed speech is given below.

The artificial language used consists of two syntactic categories. These can be thought of crudely as nouns and verbs. They can be combined to form two-word sentences. Sentences may be of the form NV or VN. Forms NN and VV are disallowed. In addition, a number of sentences are disallowed. Let us imagine that there are four nouns (n_1-n_4) and four verbs (v_1-v_4) in the language, and that v_4 is blocked in the sentence final position. From this it follows that four sentences are disallowed: each of the four nouns in combination with v_4 in an NV-type sentence.

Each model is comprised of 4 elements: word-level categories, sentence-level categories, exceptions, and code-length. Both models described here contain two word-level categories, comprising nouns and verbs and two sentence-level categories comprising the two sentence types (NV and VN). The exceptions category discretely specified all the disallowed sentences. In the first model this was an empty set. The code-length specified length of code, in bits, that would be needed to specify models just described and the corpus data given the model structure. The code-length for each sentence in the corpus is consequent on the model structure.

Calculating Code-Length For Each Element

The length of code necessary to specify any object, *i*, is given by:

$$\text{Bits}(i)=\text{Log}_2(1/p_i) \quad [1]$$

where p_i is the probability of object *i*. In many cases described below, p_i can be thought of as choosing one of I options. Where this is the case,

$$\text{Bits}(i)=\text{Log}_2 I \quad [2]$$

This section describes how this formula is applied to calculate the code-length for each section of the model and for the data given the model.

If a language contains r word types and n syntactic categories, then the probability of specifying one distribution of word types into categories is the inverse of the number of ways in which r word types can be distributed between n categories, assuming no empty sets. This given by:

$$\text{Distributions}(r, n)= \sum_{v=0}^{n} (-1)^v \frac{(n-v)^r}{(n-v)!\,v!} \quad [3]$$

The codelength for the word-level element is therefore:

$$\text{Word-level bits}(r, n)$$

$$=\text{Log}_2 \sum_{v=0}^{n} (-1)^v \frac{(n-v)^r}{(n-v)!\,v!} \quad [4]$$

Specifying a particular sentence-level rule (e.g. that a sentence may be of the form NV) is a function of the probability of that sentence type given the number of categories specified in the word-level element. Given that in the artificial language sentences only ever contain two words, there are four sentence types possible from two syntactic categories (NN, NV, VN, VV). The probability of any sentence type (e.g. NV) is therefore 1/4. When this has been specified, the probability any remaining sentence type (e.g. VN) is 1/3. The code-length for specifying two sentence types is therefore:

$$\text{Sentence-level bits}=\text{Log}_2(4)+\text{Log}_2(3) \quad [5]$$

Specifying the cost of an exception is the same as specifying the cost of a sentence. This is done by specifying the cost, in bits, of the first word based on the probability of its occurrence, and the cost of the second word in the same way. The probability of a word's occurrence is the inverse of the total number of possible words. The term to specify the first word in any sentence is therefore:

$$\text{Bits}(i1)=\text{Log}_2(\text{Tw}-T_{e1}) \quad [6]$$

where Bits(*i*1) is the bits required to specify word *i* in the first position, T_w is the total number of word types in the language and T_{e1}, is the total number of words blocked in the sentence initial position as listed in the exceptions category.

The first word specifies which sentence type is being used. The pool of possible words from which the second word must come is therefore reduced to the size of the sentence final category as defined by the sentence type. For example, if the first word in a sentence is a noun, the sentence type must be NV and the second word must therefore be from the category V. The term to specify the second word in a sentence is therefore:

$$\text{Bits}(j2)=\text{Log}_2\,(T_{wc}-T_{e2l1}) \qquad [7]$$

where Bit(j2) is the number of bits required to specify word j in the second position, T_{wc} is the total number of word types in category c, and T_{e2l1} is the total number of words specified in the exceptions element as blocked in position two given the word in position 1. The number of bits for specifying any sentence i,j is simply:

$$\text{sentence bits}_{i,j}=\text{Bits}(i1)+\text{Bits}(j2) \qquad [8]$$

Specifying the code length for each exception is the same as specifying code length for a sentence *given the existing exceptions*. Each exception in a list of exceptions therefore requires slightly fewer bits to code than its predecessor.

It is important to note that these models code corpus data in batch mode – the order in which sentences are coded is not taken into account. A more psychologically realistic (i.e. incremental) algorithm might make use of the fact that frequently occurring words have a higher probability of occurrence and therefore cost less to code.

Simulating recovery from overgeneralization with an artificial language

The models described above were implemented in a computer program. They were then exposed to successively large corpora of sentences from an artificial language which reflected the structure of the transitive/intransitive alternation phenomena found in the CHILDES database (see Table 2, above). A model using raw CHILDES data would have been computationally impossible, but it is important to note that the artificial language closely mirrored the patterns of Table 2. The artificial language is outlined above. In these simulations the word-level categories contained 36 verbs, reflecting the number of verbs in Table 2, and 36 nouns. It was decided to keep the number of nouns equal to the number of verbs in order to avoid disparity between the code-length necessary for different sentence types. There were two sentence-types (NV and VN) reflecting the transitive and intransitive contexts of the verb constructions. Ten verbs were blocked with all 36 nouns for each sentence type (see Table 2), resulting in a total of 720 disallowed sentences.

Two of the four-element models described above were exposed to increasingly large corpora of this language. The first model contained word-level information about the 36 nouns and verbs, and sentence-level information about the NV and VN sentence types, but the exceptions element was empty: it did not contain any information about the 720 disallowed sentences. In this respect it was analogous to a learner who has acquired knowledge of word categories and sentence production rules, but has not learned that some sentences are illegal. This model would therefore be prone to overgeneralizations such as *I disappeared the rabbit*. The second model, by contrast, did contain information about the disallowed sentences. This model therefore required considerably more bits to specify initially, but the number of bits required to specify each sentence of the corpus was fewer. In addition, a language learner who had learned these exceptions would not make the same overgeneralization errors that the first model would. Table 3 shows the relative simplicity of each model for increasingly large corpora as measured by the number of bits necessary to encode the model and the corpus data.

Table 3: Code-lengths of Models 1 and 2 for successively large corpora. Code-lengths in bold show the shorter codes for the corpus size

Corpus Size (sentences)	Model 1: Codelength (bytes)	Model 2: Codelength (bytes)
0	**0.1**	7.6
4000	**45.4**	51.1
8000	**90.8**	94.7
12000	**136.2**	138.3
16000	**181.5**	181.8
20000	226.9	**225.4**
24000	272.2	**268.9**

It can be seen that for relatively small corpora (up to about 16,000 sentences), Model 1 gives a simpler encoding: less bits are required. For a learner who had heard relatively few alternation constructions, therefore, the tendency would be to code the data in these terms, resulting in overgeneralizations. For a more experienced learner, however, the simpler encoding would be that shown by Model 2, which requires less bits to encode relatively large corpora. The model does not produce any language, so there are no accuracy statistics. Rather, it is assumed that the learner produces all the sentences available in the current (shortest) hypothesis are produced, including any that are incorrect.

Conclusions and future directions

These results provide an initial confirmation that simplicity may provide a guiding principle by which some aspects of language may be learned from experience without recourse to a specific language-learning device. However, the simulations presented here are coarse-grained approximations of both the language and the language learner. Children do not process the language in batches of several thousand utterances. The models presented here were neither exposed nor sensitive to different word-type frequencies. A number of further studies which would provide considerably firmer support for the simplicity principle as a driving force for language acquisition suggest themselves.

Firstly, mathematical results show that word-type frequencies are important to the simplicity-driven learner, in that they may be the key as to when it becomes advantageous to posit exceptions to rules. Chater and Vitányi (2001) show that languages are approximately learnable given sufficiently large amounts of data. The CHILDES data in Table 2 therefore provides an indication of the order in which one would expect the learner to cease overgeneralizing words. An examination of children's speech that confirmed this order would be a major step towards providing robust support for the simplicity principle in language. Secondly, it would be useful to compare the timescale of recovery from overgeneralization in children with that of the model. This could be done by an examination of CHILDES database to determine an approximate relation between a child's age and the number of transitive/intransitive alternation constructions to which they have been exposed. It would then be possible to compare the learning rate of the child with that of the model. Again, this would be a useful source of evidence concerning the simplicity principle in language.

In this paper we have suggested that there is sufficient statistical information in the input for a learner to learn quasi-regular alternating structures. These results are achieved by choosing the model of the language which provides the simplest (shortest) description of the linguistic data that has been encountered. These results re-open the question of the viability of language learning from positive evidence under less than ideal conditions, with limited computational resources and amounts of linguistic data available. They therefore also bear, indirectly, on the arguments concerning the balance between nativism and empiricism in language acquisition. More concretely, we suggest that the working hypothesis that the search for simplicity is a guiding principle in language acquisition deserves serious attention.

References:

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, 10, 533-581.

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.

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

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. (1980). *Rules and representations*. Cambridge, MA: MIT Press.

Culicover, P. (2000). *Syntactic nuts*. Oxford: Oxford University Press.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems in Information Transmission*, 1, 1-7.

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

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

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. (1987). The Competition Model. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Lawrence Erlbaum.

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

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

Pinker, S. (1995). *The language instinct*. Harmondsworth : Penguin.

Quinlan, J. R. & Rivest, R. (1989). Inferring decision trees using the minimum description length principle. *Information and Computation*, 80, 227-248.

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

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