Artif Intell Rev (2006) 26:211–225
DOI 10.1007/s10462-007-9058-x

The Simplicity and Power model for inductive inference

Emmanuel M. Pothos · J. Gerard Wolff

Published online: 21 November 2007

Abstract With this paper we wish to present a *simplicity* (informally 'simple explanations are the best') formalism that is easily and directly applicable to modeling problems in cognitive science. While simplicity has been extensively advocated as a psychologically relevant principle, a general modeling formalism has been lacking. The Simplicity and Power model (SP) is a particular simplicity-based framework, that has been supported in machine learning (Wolff, Unifying computing and cognition: the SP theory and its applications, 2006). We propose its utility in cognitive modeling. For illustration, we provide SP demonstrations of the trade-off between encoding with whole exemplars versus parts of stimuli in learning and the effect of wide versus narrow distributions in categorization. In both cases, SP computations show how simplicity can account for these contrasts, in terms of how the frequency of individual exemplars in training compares to the frequency of their constituent parts.

Keywords Simplicity · Cognitive science · Categorization · Learning

1 Introduction

When we see a series of numbers like 1...2...3...4...5... most of us assume that the sixth number is 6. In other words, we assume that the function that describes the series is $f(n) = n$, where n is the position in the series we are interested in. However, an infinite number of functions are consistent with having 1, 2, 3, 4, 5 in the first five positions, and a number *other than* 6 in the sixth position, such as $f(n) = n$ if $n < 6$, $f(n) = n^2$ if $n > 5$. Equally, when we see a circle partly obscured by a square we do not typically assume we have a strange shape that is part square part circle, but rather a circle obscured by a square

E. M. Pothos (✉)
Department of Psychology, Swansea University, Swansea SA2 8PP, UK
e-mail: e.m.pothos@swansea.ac.uk

J. G. Wolff
e-mail: jgw@cognitionresearch.org.uk
URL: CognitionResearch.org.uk

(Chater 1996; Pomerantz and Kubovy 1986; Pothos and Ward 2000). These are all applications of the *simplicity* principle.

The simplicity principle, informally Occam's Razor, states that when two hypotheses explain an observation equally well we should prefer the simpler one. More formally, if we define the complexity of a data set D as $C(D)$ and the complexity of different hypotheses to explain the data as $C(H_i)$, we should then select the hypothesis for the data to minimize the value of $C(H_i) + C(D|H_i)$, where $C(D|H_i)$ is the complexity of data D when encoded with hypothesis H_i (Solomonoff 1964, 1978). For example, both $f(n) = n$ and $f(n) = n$ if $n < 6$, $f(n) = n^2$ if $n > 5$ perfectly explain the sequence of numbers above ($C(D|H) = 0$), but the first hypothesis is (intuitively) simpler than the second. These ideas can be algorithmically implemented in several ways, one of which is the Simplicity and Power model (SP; Wolff 2006; for a general overview of related approaches see Grunwald et al. 2005). SP proposes that a given body of information should be *Simplified* by removing as much as possible of its redundant content whilst retaining as much as possible of its non-redundant descriptive *Power*; hence the name 'SP'. Of course, there have been several alternative algorithmic formulations of the simplicity intuitions, notably the Minimum Description Length (MDL; Rissanen 1978, 1987, 1989) and Minimum Message Length (MML; Wallace and Boulton 1968; Wallace and Freeman 1987; see also Baxter and Oliver 1994) approaches. Moreover, Bayesian inference (e.g., Howson and Urbach 1993) is intimately related to simplicity (Chater 1996, 1997; Chater and Manning 2006; Tenenbaum and Griffiths 2001). Elsewhere we have examined extensively the relation between SP and such approaches (Wolff 2003, 2006); a corresponding brief discussion will be presented later.

In problems of inductive inference there is no 'correct' answer. Simplicity is a guide for how to generalize from past experience to novel instances—a guide that 'instructs' the cognitive system that 6 is a reasonable continuation, whereas 1028 is not. Cognition, fundamentally, has to solve problems of inductive inference. Other prominent hypotheses for how the cognitive system makes inductive inferences are similarity (Nosofsky 1989; Vokey and Brooks 1992), rules (Hahn and Chater 1998; Pothos 2005; Reber 1989; Rips 1989), and fragments (i.e., parts of stimuli; Knowlton and Squire 1996). For example, a 'similarity' strategy for inductive inference would generalize some observed stimuli to novel ones in terms of the similarity of the novel stimuli to the old ones. It might be possible to interpret a simplicity strategy in terms of e.g., similarity or rules, but this is a complex issue beyond the scope of the present work (cf. Oaksford and Chater 1994; Pothos 2005).

Simplicity intuitions are widespread in cognitive psychology. In perceptual organization, the Gestalt view is that understanding perceptual information involves selecting the simplest interpretation of a distal layout (Chater 1996; Hochberg and McAllister 1953). Feldman (2004) used the simplicity principle to model how people decide whether there is regularity in a feature pattern, and, generally, there have been several models of figural goodness based on simplicity (Helm and Leeuwenberg 1996). In categorization, Pothos and Chater (2002) found that people often spontaneously classify a set of items in a way that simplifies the description of their similarity structure. This proposal is similar to Love et al. (2004) SUSTAIN model, whereby the creation of clusters is guided by 'surprising events.' Also, Chater and Hahn (1997) modeled psychological similarity between two objects as the algorithmic ease with which one can be transformed into the other. In learning there have been simplicity proposals of what structures develop in memory over time (e.g., Garner 1974; Miller 1958; Pothos and Bailey 1999). Simplicity has been successfully applied in reasoning as well, with Oaksford and Chater showing how our naïve intuitions on logical problems are often guided by a prerogative to minimize uncertainty in the possible conclusions (Oaksford and Chater 1994; Chater and Oaksford 1999; Oaksford et al. 1997). Note that simplicity models are

typically parameter free. Given an estimate for the complexity of data and hypotheses, the preferred hypothesis emerges automatically.

It is possible that the 'success' of the cognitive system can be partly explained by its preference for simplicity—there are formal computational reasons to prefer simplicity in inductive inference, in terms of optimal data prediction and hypothesis identification (Chater 1999; Li and Vitanyi 1997; Vitanyi and Li 1997). Is it possible that simplicity is a general principle of cognition (cf. Norenzayan and Heine 2005)? In order to begin to address this issue, we first require a framework general enough to allow simplicity predictions in a wide range of cognitive modeling situations. SP is such a candidate framework.

2 SP

SP is a theory of artificial computing and human cognition (Wolff 2003, 2006). A SP system will look to encode a new stimulus by aligning its features with the features of old stimuli (and their parts) that are stored in its memory. Also, old stimuli/parts that have been encountered frequently are associated with low codelength. The combination of old stimuli/parts to be preferred in the encoding of a new stimulus is the one that is associated with the least *overall* codelength. Hence, the better the overlap between an old stimulus and the new one, and the higher the frequency of the old stimulus, the more likely it is that this old stimulus will be preferred in the encoding of the new one (Fig. 1). This scheme (to be discussed in more detail shortly) represents a particular algorithmic interpretation of the simplicity principle. Our objective in this paper is to illustrate how SP can be readily utilized to provide predictions for psychological performance. There is a reasonable intuition that SP will be a suitable framework for modeling cognition: First, as discussed above, there is ample evidence that simplicity guides cognitive inference and generalization. Second, the SP coding assumptions are broadly consistent with the most common representational assumptions in cognitive science, whereby similarity is computed either as instance overlap (e.g., Nosofsky 1989) or feature overlap (e.g., Tversky 1977).

We next briefly examine the relation between SP and related approaches. For an extensive discussion of these issues see Wolff (2006). First, SP incorporates the key insight of algorithmic interpretations of the simplicity principle, such as MDL (Rissanen 1978) or MML (Wallace and Freeman 1987): That good hypotheses should minimize $\mathrm{C}(H + D|H)$, where $\mathrm{C}(H)$ is hypothesis complexity and $\mathrm{C}(D|H)$ is complexity of the data when encoded with H. SP deviates from such approaches by not incorporating the assumption that the Turing model is the most appropriate fundamental model for computation. Instead, SP introduces a new model of information processing, where the objective is to create *multiple alignments* that score well in terms of simplicity. In practical terms, MDL/MML is framed in terms of the sizes of the raw data (in bits) and the size of a 'program' for a Turing machine that will encode the raw data. By contrast, the SP theory makes no reference to the Turing model, it merely focuses on the sizes (in bits) of the raw data, the 'grammar', and the raw data that has been encoded in terms of the grammar. Additionally, the MDL/MML framework contains no concepts of pattern matching, heuristic search, or multiple alignment (see Sect. 4.4.1 of Wolff 2006). Second, at a general level, an equivalence between simplicity approaches and Bayesian ones can be established when the latter incorporate universal prior probability distributions, which correspond to the a priori complexity of different possible entities (Chater 1996, 1999). In general, Bayesian inference is very successful where it is possible to specify prior probability distributions, less so where there is little knowledge about the items to be processed (cf. Hines et al. in press). In many everyday life situations, the cognitive system is

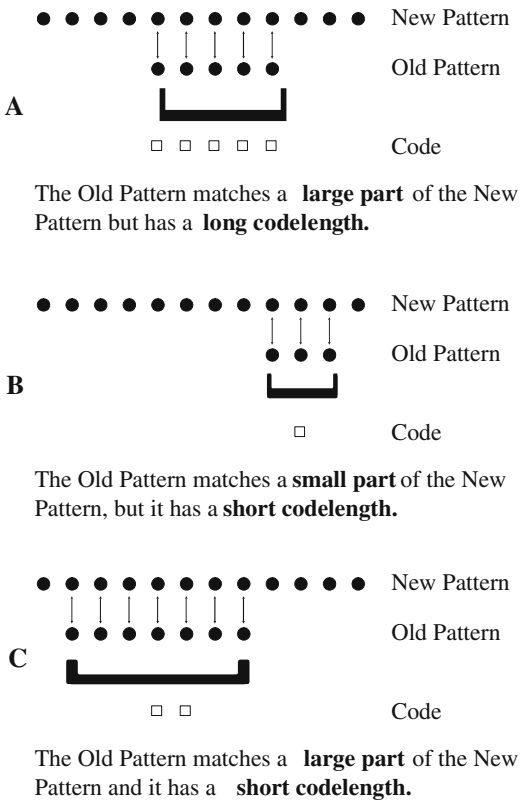

Fig. 1 The most advantageous coding of the New pattern is C since in C the Old pattern has a brief code and provides good overlap with the New pattern

asked to make inferences about objects for which there is little prior knowledge. Accordingly, while a simplicity strategy would be suitable, this is less so with a Bayesian one. Third, in associative learning stimulus encoding initially involves processing in terms of individual constituent elements; as the system gains experience, encoding gradually involves chunks of co-occurring symbols (e.g., Wasserman and Miller 1997). As is the case with SP, more frequent chunks are preferred in encoding (cf. Boucher and Dienes 2003). By contrast, in SP the encoding advantage of large chunks is concurrently considered with that of individual symbols from the outset. While this assumption deviates considerably from associative learning theory, some learning models have recently adopted it with some success (Perruchet et al. 2002). These are the key issues on how SP compares with the most directly related theory. While it should be clear that SP is closely related to MDL/MML approaches, there are important differences between SP and such approaches as well (for further discussion see Wolff 2003, 2006). Note that a direct comparison of the advantages of SP versus MDL/MML in cognitive modeling is a very complex enterprise (and certainly beyond the scope of this work). To carry out such a comparison, one would require that SP and MDL/MML approaches are applied in a series of similar domains of cognitive modeling, and subsequently compare modeling advantages. Unfortunately, this will not be possible for a while, as there have been relatively few MDL cognitive models that can lead to testable behavioral predictions (e.g., Pothos and Chater 2002, 2005; cf. Oaksford and Chater 1994).

SP can be specified in the following way. Knowledge can be represented as *patterns*, where a pattern is an array of *symbols* in one or two dimensions. Such a representational assumption has been argued not to restrict the explanatory power of cognitive models (Helm and Leeuwenberg 1996). Each symbol is considered an elementary, irreducible entity that may be compared with other symbols to decide whether the two are 'same' or 'different'. For example, in modeling linguistic processes the symbols may be letters and in categorization object features. A parameter in SP theory, the individual symbol 'cost factor,' is a measure of how 'information rich' individual symbols are (cf. Marr 1982). Increasing the cost factor of individual symbols is appropriate where the elementary symbols in the patterns are complex entities (such as entire words in the case of encoding sentences or entire stimuli in the case of modeling categorization results).

The SP model makes a distinction between New patterns and Old patterns representing stored knowledge. The objective of the system is to identify as *simple* as possible encodings for the New patterns, by matching them with Old patterns. New patterns correspond to information the cognitive system is asked to process (e.g., sensory perceptual input, a novel sentence, an object that has to be categorized) and Old patterns correspond to the system's experience. Accordingly, the process of encoding New patterns with ID-symbols can be understood as the process via which the cognitive system interprets new input, whatever its form. Similar assumptions are extremely commonplace in cognitive science (e.g., Anderson et al. 2004; Nosofsky 1989; Pothos 2005).

Each Old pattern contains an *identification* symbol (an 'ID' symbol, usually a numeral id) which serves as a name for the pattern and is relatively short compared with the size of the pattern. In accordance with Shannon–Fano–Elias coding (similar to Huffman coding), ID-symbols that are frequent are encoded with fewer bits than ID-symbols that are rare, in order to promote economical encoding of information. The number of bits required to encode an ID-symbol is calculated as: $l(x) = \log_2 \frac{1}{p(x)} + 1$, where $p(x)$ is the probability of occurrence of an ID-symbol x, derived from the frequency of occurrence of x. The intuition is exactly the same as when we shorten names of concepts that we refer to frequently, so that 'omnibus' becomes 'bus'. A New pattern may be encoded with ID-symbols from one, two, or more Old patterns that, together, provide a match for the New pattern. However, it is possible that no such suitable combination will be identified (e.g., because there are not any similar Old patterns), in which case the New pattern will be encoded as a sequence of basic symbols. (The number of bits required for encoding individual symbols is determined in a way analogous to that for ID-symbols and depends on the cost factor parameter).

These ideas are formalized with three quantities, *Compression Difference* (CD), *New Symbols Cost* (NSC), and *Encoding Cost* (EC). If Old patterns are used in the encoding of a New pattern, then EC reflects the cost associated with the ID-symbols of these Old patterns. NSC is the cost of encoding the New pattern in terms of its individual symbols, without any Old patterns. The compression difference is calculated as:

$$\mathrm{CD} = \mathrm{NSC} - \mathrm{EC}.$$

Higher CD values mean 'better' (and psychologically more intuitive) encodings. EC will be small when the corresponding Old patterns are frequent (and hence their ID-symbols encoded briefly) and when few Old patterns are used in the encoding of the New pattern. That is, EC is least when few, frequent Old patterns encode a New pattern. For Old patterns to be preferred to individual symbols, EC has to be less than the NSC, the cost to encode the New pattern with individual symbols.

Given a set of Old patterns, for any New pattern, SP generates possible alignments, each of which shows how the New pattern may be aligned with one or more Old patterns. Heuristic

search methods are used to avoid combinatorial explosion and identify alignments with high CD values (Wolff 2006). It is difficult to fully explicate the alignment process in a limited space; very briefly, it involves the following steps (for full details and examples please see Wolff 2006):

1. Initially, the model looks for 'good' alignments, where each alignment is between the New pattern, or part of the New pattern, and any one of the Old patterns. The process of searching for 'good' alignments of two patterns is an enhanced version of the concept of 'dynamic programming' (see Sect. 3.10.3.1 and Appendix A in Wolff 2006).
2. Any alignments containing mismatches are rejected (as described in Sect. 3.4.7 of Wolff 2006).
3. Each of the remaining alignments can be 'unified' and treated as if it was a single sequence of symbols (although the model remembers how each alignment was built up; note, generally, that each multiple alignment shows exactly which Old symbols match which New symbols). The model takes the best of these partial alignments and looks for ways to align these partial alignments with any of the Old patterns or with each other.
4. The process returns to step 2 and keeps repeating steps 2 and 3 until no more 'good' alignments can be found. From the best alignment it is possible to derive an economical encoding for the New pattern.

Approximately 30 parameters control the operation of the software implementation and have no explanatory content—most relate to either the format of the output or the extent of the search. Parameters that have 'explanatory' content (in other words, the parameters which may affect the modeling results) are the cost factor of individual symbols and two parameters that determine whether all symbols in the New/Old patterns used in an alignment have to be matched. Whether all the symbols of New/Old patterns have to be matched or not is a choice that depends partly on the modeling situation. For example, in encoding (understanding) a sentence in terms of words (and combinations of words) from experience, it is not a requirement to encode every single word of a new sentence. In other words, we can still understand a sentence even if not every single one word has been encoded. All SP parameters are set a priori. A C++ implementation of SP (where a list of the parameters can be found) may be downloaded from: http://www.cognitionresearch.org.uk/source_code/SP62.ZIP. Our purpose in this section was to provide a description of the basic principles underlying the SP formalism. It is beyond the scope of this work to consider all the algorithmic details of the SP model—an extensive exposition is Wolff (2006).[1]

In the subsequent sections we examine two applications of SP, relating to two extensively studied paradigms in cognitive science, artificial grammar learning and categorization. We show that the SP application can lead to novel insights about these tasks and an understanding of their relation.

[1] Note that here lies the ability of the SP model to learn: Briefly, every Old pattern contains symbols of two kinds: C-symbols representing the 'contents' of the pattern and ID-symbols representing a 'code' or 'identifier' for that content. If the requirement that all elements of the New pattern are matched is in place, the ID-symbols of an Old pattern can effectively be used to encode the New pattern. Whenever a partial match between a New pattern and the C-symbols of an Old pattern is allowed and exists, this is a signal that learning should occur. Here, learning means that the system creates new patterns (each one containing C-symbols and ID-symbols), and it does it in such a way that, after learning has occurred, it is possible to create a multiple alignment in which all the C-symbols in all the patterns of the alignment have been matched (for more details see Wolff 2006). The present simulations employed a simpler version of SP that does not learn (that is, the system's experience already contains all the necessary Old patterns).

3 Fragments versus individual exemplars

Artificial Grammar Learning (AGL) is a widely used paradigm for the study of implicit learning (Reber 1989). AGL involves exposing participants to a set of stimuli in a training phase and subsequently asking participants to decide which new stimuli are of the same kind as the original ones in a test phase. The stimuli in an AGL task are generated by a set of rules (a finite state language), so that some of the test stimuli will comply to the rules (grammatical, G), while some will violate them (nongrammatical, NG). It has been shown that the test items preferred as G are the ones that are more similar to whole training instances (Pothos and Bailey 2000; Vokey and Brooks 1992) or the ones that are made of fragments that are frequent in training (Knowlton and Squire 1996; other AGL hypotheses are not presently relevant). Meulemans and van der Linden (1997) showed that when there are few training items, similarity effects are enhanced, but there has been no proposal of a computational account to predict when whole exemplar similarity will be preferred to fragments.

A similar contrast arises in categorization, with exemplar and prototype theories of categorization. In exemplar theories a new instance X is more likely to be classified as a member of the category A, if X is more similar to each of A's members (Nosofsky 1989). In prototype theories, classification of A is determined by its comparison to a summary representation of A's members (its prototype; Posner and Keele 1968). Some investigators have argued that increasing individual exemplar salience (e.g., by increasing exemplar frequency or distinguishability) enhances the likelihood of whole-exemplar classification (Regehr and Brooks 1993; Rouder and Ratcliff in press; Shin and Nosofsky 1992; cf. Pothos 2005). These results can be explained in the Generalized Context Model of categorization, in terms of increased 'strength' for the salient exemplars (Shin and Nosofsky 1992), or by variants of Logan's ideas on the development of automaticity (Logan 1988; Rouder and Ratcliff in press). Logan suggested that the more experience we have with a thematic domain, the more likely it is to employ whole exemplars in processing new situations, as opposed to trying to work out the situations in some algorithmic way. For example, in starting to learn how to ride a bicycle, the learner often has to consciously consider the sequence of actions which needs be performed. By contrast, according to Logan, experienced bicycle riders automatically activate previous instances of 'cycling' experiences when they need to ride a bicycle. In other words, this in another way in which increasing the salience of certain exemplars in the experience of a person will make it more likely for these exemplars to be used in processing new information.

Encoding in terms of prototypes can be seen as similar to encoding in terms of fragments: In both cases, there is a sense in which new instances are encoded by a summary representation of the instances in our experience. Hence, we suggest that there is an analogy between the prototype/exemplars distinction and the fragments/exemplars one (of course, superficially, prototypes are very different from frequent fragments).

We will show that in AGL an explanation of when fragments will be preferred to whole exemplars can be understood in terms of simplicity/efficiency of encoding considerations. Additionally, SP predicts that generalization will be sensitive to small changes in exemplar frequency, but changes in symbol/fragment frequencies will have to be considerable before they have any effect.

We first ran two simulations, with the same New pattern and the same Old patterns. SPs task was to encode (match) a single New Pattern in terms of a set of Old Patterns in Table 1. In one simulation (exemplar case) the frequency of individual exemplars and individual symbols was increased; in the other (fragments case) the frequency of the bigrams (pairs of letters)

Table 1 Note even individual symbols need be explicitly represented as Old patterns

Exemplar case	Fragments case
1 A *100	1 A
2 B *100	2 B
3 C *100	3 C
4 A B	4 A B *50
5 A C	5 A C *50
6 B C	6 B C *50
7 A C	7 A C *50
8 A A	8 A A *50
9 B B	9 B B *50
10 C C	10 C C *50
11 A A A B C *10	11 A A A B C
12 A B A B A *10	12 A B A B A

The number in front of each Old pattern is its id. 'A * 100' means that the frequency of Old pattern A is 100

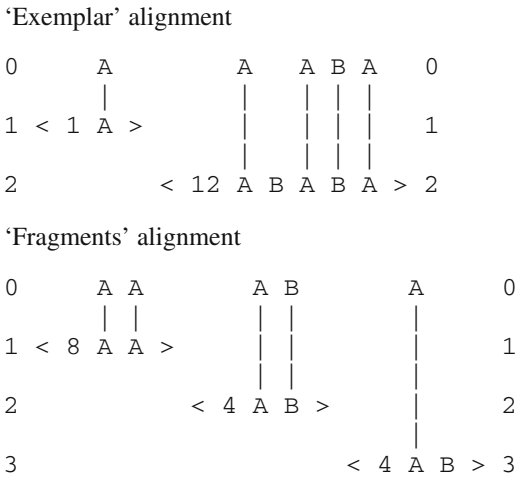

Fig. 2 Alignments for the New pattern AAABA on the basis of the two Old pattern sets in Table 1. The New pattern is shown in the first row and in each of the subsequent rows each of the Old patterns used in the alignment are shown. The id symbol of each Old pattern is shown in front of it—it is on the basis of these id symbols that the EC for the alignment is computed

making up the exemplars was increased (see Table 1). The idea was to examine when SP will prefer encodings with individual exemplars (and symbols), as opposed to fragments. Note that in cognitive processing there seem to be situations where our experience with symbols/bigrams is not limited to our experience with corresponding whole exemplars and vice versa (e.g., letters of the English alphabet).

Cost factor was set to 150 in all simulations (here and elsewhere), New pattern symbols had to be all matched, symbols of Old patterns used in the alignment did not (several discussions of SP simulations can be found in Wolff 2006). The best alignments in the exemplar and fragments case are shown in Fig. 2, using SP notation. As can be seen, in the exemplar case a whole Old pattern similar to the New one is preferred in the encoding, but in the fragments case several Old pattern bigrams are employed. SP generally prefers alignments

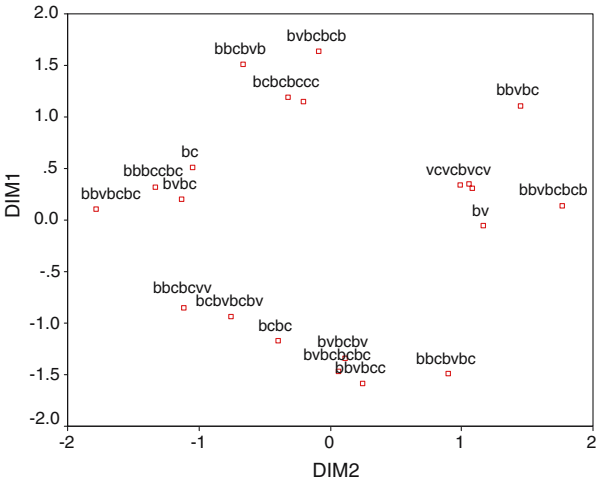

Fig. 3 An MDS representation of the similarity structure of the Patterns used in the 'single exemplar effects' demonstration. The New pattern is bbvbc. Some of the item labels have been removed for clarity

involving entire exemplars and the frequency of bigrams needs be increased a lot more than the frequency of specific exemplars before they are preferred in an alignment.

We develop this intuition by using a set of letter strings. Appendix 1 lists the items, and their trigrams/bigrams/symbols and corresponding frequencies, and the single New pattern. The Old patterns category structure can be explored as follows: First, letter string similarity is computed in a way that is relevant to SP, as fragment overlap between any two strings. By analogy with Knowlton and Squire (1996), we defined the similarity between two strings as the proportion of bigrams/trigrams common between the strings, relative to the maximum number of fragments that could be common, from left to right, retaining positional information (the maximum possible number of common chunks is the number of chunks in the longer of the two strings that are compared). In this way we built a similarity matrix between all stimuli. For illustration, a spatial representation of the letter strings could then be derived using Multidimensional Scaling (MDS; stress: 0.40, Fig. 3).

A random New pattern was chosen ('bbvbc'). As before, SP was required to match all symbols in the New pattern and we explored alignments with and without the requirement that all symbols in the Old patterns used in the alignment are matched. Three simulations were ran. In Simulation 1 the fragment and symbol frequencies were as shown in Appendix 1. The best alignment involving fragments was made with 'bbv' and 'bc' (CD = 1973), and involving a single exemplar was made with 'bvbvbvbc,' (CD = 2005). Other single exemplars were involved in alignments that had almost identical CD values. In Simulation 2, the frequency of symbols and fragments was increased by a factor of 10. The best alignment involving fragments (as before, 'bbv' and 'bc'; CD = 2075.61) had a CD value nearly identical to the best alignment involving a single exemplar (as before, 'bvbvbvbc'; CD = 2075.91). In Simulation 3, the frequency of the fragments was as in Appendix 1 but we increased by 2 the frequency of one of the Old patterns that led to the best CD value in Simulation 1. The best alignment now corresponded to this Old pattern.

In sum, SP prefers alignments involving single exemplars; alignments in terms of parts will be preferred when the frequency of the parts is very high. SP allows a formulation of these intuitions in a way that can lead to quantitative predictions.

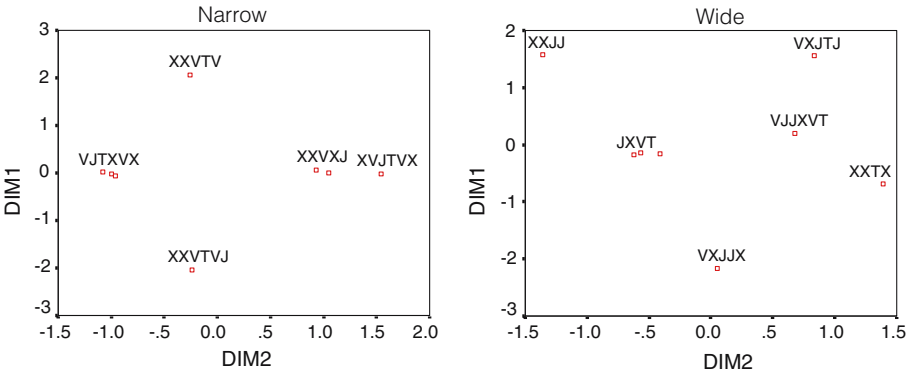

Fig. 4 An MDS representation of the similarity structure of the Patterns used in the 'wide versus narrow distributions' demonstration. Some of the item labels have been removed for clarity

4 Wide versus narrow distributions

We presently generalize the previous demonstration and examine whether categories with different distributional characteristics can lead to specific similarity or fragments-based generalization (this will lead us to a conclusion regarding exemplar vs. prototype theories of categorization). Since Rips' (1989) pizza-coin demonstration, it has been appreciated that more variable categories will allow more flexible extensions to novel exemplars (Hahn et al. 2005; Mareschal et al. 2002; Smith and Sloman 1994). Hahn et al. explained this result in terms of participants becoming less sensitive to dimensions along which the members of a category are more variable and Mareschal et al. by assuming that learning a less variable category after a more variable one effectively reinforces many of the representations already in place from having learned the more variable category. We explore SP generalization from categories with narrow and wide distributions to find that in the former case generalization involves individual whole exemplars and in the latter fragments. Hence, this demonstration shows that the category variability effect (a) has an interpretation in terms of simplicity/encoding efficiency considerations (b) it effectively reflects the same contrast between encoding in terms of exemplars versus fragments, as discussed in the previous section.

The AGL test stimuli of Knowlton and Squire (1996, Experiment 1) were created to be members of one of four categories: G and high similarity to the training items, G low similarity, NG high similarity, and NG low similarity. Hence, legality of the strings and similarity to the training items were counterbalanced. Similarity of test items to the training ones was assessed as a function of fragment overlap. What is presently relevant is that the G high similarity items form a category with a narrower distribution, compared to the NG low similarity ones—NG items are still somewhat coherent, but since they are created to be both illegal and dissimilar to the training items, they are typically more variable. We confirmed this expectation independently.

We defined the 'narrow' distribution category to consist of the eight G high similarity items of Knowlton and Squire (1996, Experiment 1) and the 'wide' distribution category to consist of the eight NG low similarity items (Appendix 2). We then computed the number of *unique* bigrams/trigrams in all the items of each category, as a measure of category variability. For the narrow category there were 21 unique fragments and for the wide one 29.

Table 2 Number of Fragments employed in the cases of categories with narrow, wide distribution of exemplars

Novel pattern	Best alignment involved patterns	CD value	Number of fragments involved
Narrow category			
VXJJJJ	VXJ, J, J, J	3285	1
VJTVXJ	VJT, VXJ	3078	2
XXVXJJ	XXVXJ, J	2915	0
XXXVTV	X, XXVTV	2639	0
VJTXVJ	VJT, XVJ	3078	2
XVJTVJ	XVJ, TVJ	3077	2
XVXJJJ	X, VXJ, J, J	3106	1
XXXXVX	XX, XXV, X	2444	2
VJTVTV	VJT, VTV	3005	2
Wide Category			
VXJJJJ	VXJ, JJ, J	2628	2
VJTVXJ	VJ, T, VXJ	2947	2
XXVXJJ	XXV, XJJ	2657	2
XXXVTV	XX, XVT, V	2947	2
VJTXVJ	VJ, TX, VJ	2945	3
XVJTVJ	XV, J, TVJ	2947	2
XVXJJJ	X, VXJ, JJ	2647	2
XXXXVX	XX, XXV, X	2626	2
VJTVTV	VJ, TV, TV	3205	3

Using the same method as in the previous section, for illustration, we derived a spatial representation of the items in the narrow and wide categories (stress: 0.37 and 0.30, respectively; Fig. 4).

The Novel patterns were the *training* items of Knowlton and Squire (1996, Experiment 1) that were composed of six symbols (Appendix 2). So in our demonstration some of the training and test items of Knowlton and Squire were used the other way round, so as to achieve the narrow/wide category design. Note that we are not modeling any result from Knowlton and Squire (1996), rather utilize their stimuli to consider the empirical issue outlined at the beginning of this section. Requiring SP to match all symbols of the New patterns and all symbols of the Old patterns used in the alignment, we identified the best alignment for each of the nine Novel patterns. For each of these alignments we computed the number of fragments that were used in the alignment: In other words, how many bigrams or trigrams were employed in the alignment? Note that the CD values for the narrow and wide categorizations are not comparable, since (trivially) the Old patterns are different in the two cases. Note also that such a scheme somewhat inflates the fragment generalization counts since some of the Old patterns are actually bigrams or trigrams; but, given the preliminary nature of our demonstration this consideration was deemed irrelevant.

Our results are shown in Table 4. Each of the nine New patterns was categorized twice, once with the narrow category, and a second time with the wide one. Alignments with the wide category involved more fragments than alignments with the narrow category in six out of nine cases; for the remaining three New patterns, narrow and wide categorizations involved alignments with the same number of fragments. In sum, classifying New patterns with the wide category stimuli consistently involved alignments with more fragments, compared to the narrow category classifications.

5 Conclusions

SP is a model of how to generalize from old experience to novel instances. Much of cognitive science deals with this problem and prominent hypotheses relate to rules, similarity, or associative learning (Pothos 2005). The relationship between simplicity approaches and such alternatives is a complex issue. For example, consider the distinction between whole exemplars and fragments, which we considered in the context of AGL, and the one between exemplars and prototypes, which we considered in categorization. In both cases, we aimed to provide an account of such distinctions on the basis of (effectively) the simplicity principle. This is not to say that a simplicity model is to replace such models as those making reference to prototypes, exemplars etc. Notions such as prototypes and exemplars have an important descriptive value when it comes to understanding human behavior. Rather, the simplicity approach is aimed to supplement such models in a way which addresses some of their current shortcomings. For example, existing cognitive psychology theory makes no predictions as to when a prototype or an exemplar mode of categorization should be employed. The simplicity approach we outlined aims exactly to address such shortcomings, rather than provide a view of cognitive processes which is mutually exclusive relative to existing theory.

Why explore simplicity? Because there are normative justifications as to why simplicity inference is computationally advantageous (Chater 1999; see also Oaksford and Chater 1991). Extensive research has shown the relevance of simplicity in cognition. SP is a theory that develops the simplicity intuitions in a computational framework appropriate for modeling a diverse range of cognitive processes.

SP was illustrated by modeling the contrast between encoding on the basis of whole stimuli versus parts of stimuli and the effects of narrow versus wide category distributions in categorization. The first issue has been prominent in learning (e.g., Knowlton and Squire 1996; Reber 1989) and categorization (e.g., Shin and Nosofsky 1992). The second one is an important finding also in categorization, and has informed theoretical progress in both the rules versus similarity distinction (Rips 1989) and the acquisition of concepts in developmental psychology (Mareschal et al. 2002). By modeling the whole exemplars versus fragments contrast in SP we showed this to be a contrast that can be interpreted in terms of encoding efficiency: Fragments or whole exemplars will be preferred depending on how *simple* encodings they provide for a new stimulus. (As said, this conclusion does not contrast with existing theory, rather it provides an additional, information theoretic, perspective). Moreover, SP computations showed that differences in generalization from narrow versus wide categories reflect exactly a difference in encoding on the basis of whole stimuli versus parts of stimuli. These are findings that (a) further inform existing theory in learning and categorization (b) allow corresponding predictions of human performance on the basis of few parameters and (c) enable an interpretation of effects, historically considered separate, in the same (SP) framework.

Clearly, the demonstrations we report do not allow us to conclude that the SP formalism is a good model for cognitive processes *in general* (Wolff 2006, discusses extensively the extent to which the SP model is good model of artificial intelligence). Our objective was to present the SP formalism as a plausible framework for modeling psychological processes (with good motivation) and demonstrate that it allows some novel insight when it comes to two specific areas of controversy in cognitive psychology.

Appendix 1

The items employed in the demonstration in the Sect. 3.

Whole exemplars		Parts	
bbbccbc	Old Pattern	c *38	Old Pattern
bbcbcvv	...	b *64	...
bbcbvbc	...	v *26	...
bbcbvb	...	cc *4	...
bbvbc	...	cb *18	...
bbvbcbc	...	cv *4	...
bbvbcbcb	...	bc *31	...
bbvbcc	...	bb *9	...
bc	...	bv *20	...
bcbc	...	vc *3	...
bcbcbccc	...	vb *17	...
bcbvbcbv	...	vv *1	...
bcvbc	...	ccc *1	...
bv	...	ccb *1	...
bbvbc	...	cbc *10	...
bvbcbcb	...	cbv *6	...
bvbcbcbc	...	cvc *1	...
bvbcbv	...	cvb *1	...
bvbvb	...	cvv *1	...
bvbvbvbc	...	bcc *3	...
vcvcbvcv	...	bcb *16	...
bbvbc	New Pattern	bcv *2	...
		bbc *4	...
		bbb *1	...
		bbv *4	...
		bvc *1	...
		bvb *16	...
		vcb *1	...
		vcv *2	...
		vbc *12	...
		vbv *3	...

Appendix 2

The items employed in the demonstration in the Sect. 4.

Narrow category	Wide category	Novel instances
XXVXJ	XXJJ	VXJJJJ
XVTV	VXJTJ	VJTVXJ
VXJ	XXVVJJ	XXVXJJ
XXVTV	JXVT	XXXVTV
XVJTVX	XXTX	VJTXVJ
XXVTVJ	TVJ	XVJTVJ
VJTXVX	VXJJX	XVXJJJ
VX	VJJXVT	XXXXVX
		VJTVTV

Acknowledgements We would like to thank Brad Love and Nick Chater for their many helpful comments and Neil Carter for his assistance with Perl programming. This work was supported in part by EC Framework 6 contract 516542 (NEST) and ESRC grant R000222655.

References

Anderson JR, Bothell D, Byrne MD, Douglass S, Lebiere C, Qin Y (2004) An integrated theory of the mind. Psychol Rev 111:1036–1060

Baxter R, Oliver J (1994) MDL and MML: similarities and differences. *Tech Report* 207, Department of Computer Science, Monash University

Boucher L, Dienes Z (2003) Two ways of learning associations. Cogn Sci 27:807–842

Chater N (1996) Reconciling simplicity and likelihood principles in perceptual organization. Psychol Rev 103:566–591

Chater N (1997) Simplicity and the mind. The psychologist. November 1997:495–498

Chater N (1999) The search for simplicity: a fundamental cognitive principle? Quart J Exp Psychol 52A: 273–302

Chater N, Hahn U (1997) Representational distortion, similarity and the Universal Law of generalization. In: Proceedings of the similarity and categorization workshop 97. University of Edinburgh, pp 31–36

Chater N, Manning CD (2006) Probabilistic models of language processing and acquisition. Trends Cogn Sci 10:335–344

Chater N, Oaksford M (1999) The probability heuristics model of syllogistic reasoning. Cogn Psychol 38: 191–258

Feldman J (2004) How surprising is a simple pattern? Quantifying "Eureka!". Cognition 93:199–224

Garner WR (1974) The processing of information and structure. LEA, Potomac, Md

Grunwald PD, Myung J, Pitt MA (eds) (2005) Advances in minimum description length: theory and applications. MIT Press Cambridge,

Hahn U, Chater N (1998) Similarity and rules: distinct? exhaustive? empirically distinguishable? Cognition 65:197–230

Hahn U, Bailey TM, Elvin LBC (2005) Effects of category diversity on learning, memory, and generalization. Mem Cogn 33:289–302

Hines P, Pothos EM, Chater N (in press) A non-parametric approach to simplicity clustering. Appl Artif Intell

Hochberg JE, McAlister E (1953) A quantitative approach to figural goodness. J Exp Psychol 46:361–364

Howson C, Urbach P (1993) Scientific reasoning: the Bayesian approach. Open Court, Chicago

Knowlton BJ, Squire LR (1996) Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. J Exp Psychol: Learn, Mem Cogn 22:169–181

Li M, Vitanyi P (1997) An introduction to Kolmogorov complexity and its applications, 2nd edn. Springer-Verlag, Berlin

Logan GD (1988) Toward an instance theory of automatization. Psychol Rev 95:492–527

Love BC, Medin DL, Gureckis TM (2004) SUSTAIN: a network model of category learning. Psychol Rev 111:309–332

Mareschal D, Quinn PC, French RM (2002) Asymmetric interference in 3- to 4-month-olds' sequential category learning. Cogn Sci 26:377–389

Marr D (1982) Vision: a computational investigation into the human representation and processing of visual information. W. H. Freeman, San Francisco

Meulemans T, van der Linden M (1997) Associative chunk strength in artificial grammar learning. J Exp Psychol: Learn, Mem, Cogn 23:1007–1028

Miller GA (1958) Free recall of redundant strings of letters. J Exp Psychol 56:485–491

Norenzayan A, Heine SJ (2005) Psychological universals: what are they and how can we know? Psychol Bull 131:763–784

Nosofsky RM (1989) Further tests of an exemplar-similarity approach to relating identification and categorization. J Exp Psychol: Percept Psychophys 45:279–290

Oaksford M, Chater N (1991) Against logicist cognitive science. Mind Lang 6:1–38

Oaksford M, Chater N (1994) A rational analysis of the selection task as optimal data selection. Psychol Rev 101:608–631

Oaksford M, Chater N, Grainger B, Larkin J (1997) Optimal data selection in the reduced array selection task (RAST). J Exp Psychol: Learn, Mem, Cogn 23:441–458

Perruchet P, Vinter A, Pacteau C, Gallego J (2002) The formation of structurally relevant units in artificial grammar learning. Quart J Exp Psychol 55A:485–503

Pomerantz JR, Kubovy M (1986) Theoretical approaches to perceptual organization: simplicity and likelihood principles. In: Boff KR, Kaufman L, Thomas JP (eds) Handbook of perception and human performance. vol. II. Cognitive processes and performance, Wiley, New York, pp 1–45

Posner MI, Keele SW (1968) On the genesis of abstract ideas. J Exp Psychol 77:353–363

Pothos EM (2005) The rules versus similarity distinction. Behav Brain Sci 28:1–49

Pothos EM, Bailey TM (1999) An entropy model of artificial grammar learning. In: Proceedings of the twenty-first annual conference of the cognitive science society, LEA, Mahwah, pp 549–554
Pothos EM, Bailey TM (2000) The importance of similarity in artificial grammar learning. J Exp Psychol: Learn, Mem, Cogn 26:847–862
Pothos EM, Chater N (2002) A simplicity principle in unsupervised human categorization. Cogn Sci 26: 303–343
Pothos EM, Chater N (2005) Unsupervised categorization and category learning. Quart J Exp Psychol 58A:733–752
Pothos EM, Ward R (2000) Symmetry, repetition, and figural goodness: an investigation of the weight of evidence theory. Cognition 75:B65–B78
Reber AS (1989) Implicit learning and tacit knowledge. J Exp Psychol: General 118:219–235
Regehr G, Brooks LR (1993) Perceptual manifestations of an analytic structure: the priority of holistic individuation. J Exp Psychol: General 122:92–114
Rips LJ (1989) Similarity, typicality and categorization. In: Vosniadou S, Ortony A (eds) Similarity and analogical reasoning. Cambridge University Press, Cambridge
Rissanen J (1978) Modeling by shortest data description. Automatica 14:465–471
Rissanen J (1987) Stochastic complexity. J Royal Stat Soc Ser B 49:223–239
Rissanen J (1989) Stochastic complexity and statistical inquiry. World Scientific, Singapore
Rouder JN, Ratcliff R (2006) Comparing exemplar- and rule-based theories of categorization. Curr Direction Psychol Sci 15:9–13
Shin HJ, Nosofsky RM (1992) Similarity-scaling studies of "dot-pattern" classification and recognition. J Exp Psychol: General 121:278–304
Smith EE, Sloman SA (1994) Similarity—vs. rule—based categorization. Mem Cogn 22:377–386
Solomonoff RJ (1964) A formal theory of inductive inference. Parts I and II. Inf Control 7:1–22, 224–254
Solomonoff RJ (1978) Complexity-based induction systems: comparisons and convergence theorems. IEEE Trans Inf Theory 24:422–432
Tenenbaum J, Griffiths TL (2001) Generalization, similarity, and Bayesian inference. Behav Brain Sci 24: 629–641
Tversky A (1977) Features of similarity. Psychol Rev 84:327–352
Vitanyi PMB, Li M (1997) On prediction by data compression. In: Proceedings of 9th European conference on machine learning, lecture notes in artificial intelligence, vol 1224. Springer-Verlag, Heidelberg, pp 14–30
van der Helm PA, Leeuwenberg LJ (1996) Goodness of visual regularities: A nontransformational approach. Psychol Rev 103:429–456
Vokey JR, Brooks LR (1992) Salience of item knowledge in learning artificial grammar. J Exp Psychol: Learn, Mem Cogn 20:328–344
Wallace CS, Boulton DM (1968) An information measure for classification. Comp J 11:185–195
Wallace CS, Freeman PR (1987) Estimation and inference by compact coding. J R Stat Soc, Ser B 49:240–251
Wasserman EA, Miller RR (1997) What's elementary about associative learning? Ann Rev Psychol 48: 573–607
Wolff JG (2003) Information compression by multiple alignment, unification and search as a unifying principle in computing and cognition. Artif Intell Rev 19(3):193–230
Wolff JG (2006) Unifying computing and cognition: the SP theory and its applications. CognitionResearch.org.uk, Menai Bridge. ISBN 0–9550726-0-3 (Ebook edition distributed by Amazon.com