J. Linguistics **41** (2005), 513–569. © 2005 Cambridge University Press
doi:10.1017/S0022226705003439 Printed in the United Kingdom

The Subset Principle in syntax: costs of compliance[1]

JANET DEAN FODOR

The Graduate Center, City University of New York

WILLIAM GREGORY SAKAS

Hunter College and The Graduate Center
City University of New York

(Received 2 July 2004; revised 16 June 2005)

Following Hale & Reiss' paper on the Subset Principle (SP) in phonology, we draw attention here to some unsolved problems in the application of SP to syntax acquisition. While noting connections to formal results in computational linguistics, our focus is on how SP could be implemented in a way that is both linguistically well-grounded and psychologically feasible. We concentrate on incremental learning (with no memory for past inputs), which is now widely assumed in psycholinguistics. However, in investigating its interactions with SP, we uncover the rather startling fact that incremental learning and SP are incompatible, given other standard assumptions. We set out some ideas for ways in which they might be reconciled. Some seem more promising than others, but all appear to carry severe costs in terms of computational load, learning speed or memory resources. The penalty for disobeying SP has long been understood. In future language acquisition research it will be important to address the costs of obeying SP.

1. Introduction

1.1 *Background*

In a previous issue of this journal, Mark Hale and Charles Reiss noted some problems concerning the application of the Subset Principle to phonological acquisition (Hale & Reiss 2003). Here, we draw attention to some problems in the application of the Subset Principle (SP) to syntax acquisition. Some of these points can be found already in the literature, but others are less familiar. We think it is useful to collate them here as a ready reference for linguists and psycholinguists whose theories presuppose some version of SP.

[1] For their helpful advice and feedback we are grateful to two *JL* referees, and the audience at the 2004 Midwest Computational Linguistics Colloquium at the University of Indiana. This research was supported in part by grants 65398-00-34, 66443-00-35 and 66680-00-35 from the Professional Staff Congress of the City University of New York.

Our message is not that all such research should come to a standstill until SP-related issues have been resolved. That would be a drastic policy indeed, when it is still not known how or even whether they are resolvable. Nevertheless, it is salutary to remind oneself every now and then, when invoking SP, that there is as yet no satisfactory theory of what work it should do, or how, even in principle, it could do it.

We will point toward solutions where we can, but in the end we leave many questions unresolved. Some of the points we make here overlap with those of Hale & Reiss, but most do not. We share with them the view that it is healthy for SP issues to be aired but we will not review their paper in detail nor try to apply our own discussion to the acquisition of phonology. Our ultimate aim, like that of Hale & Reiss, is an acquisition model that is both linguistically grounded and psychologically plausible. Psychological considerations (such as memory and computational resources) are therefore relevant in evaluating proposed learning procedures. For reasons of space we give these criteria precedence here over formal specification of a learning algorithm, though the latter is also important. The six conditions on a theory of language learning articulated by Pinker (1979) remain valid: the learnability condition must be met but so also must constraints on the time-course of learning, the nature of input, cognitive resources and so forth.

1.2 *Starting assumptions*

To prevent the discussion from spreading too far afield we must make some working assumptions that will focus attention on central issues. Some of these (A1 and A2 below) refer to properties of the initial or attained grammar; some (A3, A4, A6, A7, A8) to the strategies of the learner; some (A5, A9) to the capacities of the learner; and some (A10, A11, A12) to the nature of the language sample the learner is exposed to. Our working assumptions (except where specifically repealed) will be:

A1. that human infants have innate knowledge of all universal aspects of natural language syntax (*Universal Grammar*, *UG*);
A2. that grammars differ from each other only with respect to the lexicon, which includes a finite number of UG-defined *parameters* that control aspects of sentence structure;
A3. that learners hypothesize only grammars compatible with UG;
A4. that learners acquire novel syntactic facts only from sentences whose lexical items they are already acquainted with;
A5. that the learning mechanism is *memoryless* in the sense that at any step it has no recall of its prior input or its prior grammar hypotheses; it knows only the current input sentence and its grammar hypothesis immediately prior to receiving that input;

A6. that learning is *incremental*, in the sense that the learner hypothesizes a grammar (perhaps partial, perhaps unchanged from the preceding one) after each encounter with an input sentence; as point A5 implies, target language sentences cannot be accumulated in memory and subsequently compared in order to extract generalizations;

A7. that the learner is *error-driven*, i.e., it doesn't change its hypothesis unless it encounters input that contradicts it;

A8. that the learner is *greedy*, i.e., it does not give up its current grammar hypothesis in favor of a grammar that is incompatible with the current input sentence;

A9. that the learner can 'decode' the input, i.e., the learner knows which grammars/languages are compatible with the current input sentence (though the feasibility of such knowledge will come under scrutiny later in the paper);

A10. that the learner's input is 'noise'-free, i.e., is not contaminated by ungrammatical sentences;[2]

A11. that the input contains no direct negative evidence (i.e., no information about which word strings are ungrammatical in the target language) and the learner has no access to indirect negative evidence or other mechanisms for retreating from overgeneral hypotheses;

A12. that the input is a 'fair' presentation (sometimes referred to as a *fat text*) in the sense that no string is systematically withheld and strings may freely repeat.

These assumptions are all more or less familiar in the literature. Listing them here does not mean that we (or others) endorse them all. Their role is to facilitate understanding of the 'logical problem of language acquisition' (Roeper & Williams 1987) by revealing the impact of various aspects of the learning situation on the possibility of a successful outcome. Broad assumptions such as these inevitably present extreme oversimplifications of real-life language acquisition. Some may exaggerate how easy it is (e.g., A9, A10), while others may exaggerate its difficulty (e.g., A5, A11). Nevertheless, they provide a useful starting point for an investigation of the range of possible acquisition mechanisms. It is likely that the right way to solve SP problems will be to modify or abandon one or more of these assumptions. We will consider some possible amendments as the discussion proceeds,

[2] Language input to children may be noisy not just because of speech errors or admixture of other languages, but also if the learner misanalyzes a well-formed input. If learners cannot distinguish noisy data from genuine target language sentences, noise can undermine any attempt to prevent fatal overgeneration errors, however assiduously SP is applied. A non-target-language sentence may appear to justify a grammar hypothesis which is in fact an SP violation with respect to the target language. Like many other discussions of natural language learnability, we leave the noise problem unsolved here, though in other work we have begun to study the impact (good as well as bad) of noise on learning in a natural-language-like domain (Crowther et al. 2004).

and especially in section 4, where revisions to assumption A5 will be explored.

1.3 *The role of SP*

When a *learning mechanism* (*LM*) as characterized above is confronted with an exemplar of the target language that is not licensed by its current grammar hypothesis, it will seek to adjust the current grammar so that it does accommodate this new input. It is evident that LM at least sometimes hypothesizes a grammar which adds not only this one current input sentence but other sentences as well. This must be so, since an infinite language cannot be learned from a finite sample without generalizing. The challenge for the learner is to decide how far to generalize, and along what dimensions.

For instance, on observing topicalization of the object in the English sentence *Tom, I like*, should the learner generalize topicalization to all and only proper names? all and only nouns? singular NPs? all NPs? all XPs? On hearing an instance of an auxiliary verb preceding *not*, should the learner generalize only to other instances of the same auxiliary verb? or to all auxiliary verbs? or to all verbs? The correct answer to the first question is: all XPs (subject to some constraints!). The correct answer to the second question is: only auxiliary verbs. These alternative grammar hypotheses differ in breadth: some license languages that properly include (are proper supersets of) the languages licensed by others.[3] Choosing a broad enough generalization is important in order to move swiftly toward the target grammar. As we will discuss in section 3.2, a cautious learner may converge on the target grammar very slowly or not at all. But choosing too broad a generalization can be fatal, since by working assumption A11, LM lacks sufficient negative data to guide retreat from an incorrect superset choice. Though children do sometimes misgeneralize (which provides data of great interest), it seems that by and large they are not greatly distracted by the vast number of potential incorrect generalizations, but succeed in homing in on the correct generalization with considerable efficiency. In particular, there is little evidence that learners get trapped in superset hypotheses. If that were a typical learning problem, languages could be expected to show broader and broader generalizations as time goes on, but this is not a standardly reported finding; many language-specific limits survive through generation after generation of

[3] Two notes on terminology. (i) Throughout this paper, we use the term *subset* to mean *proper subset*, and *superset* to mean *proper superset*, unless otherwise indicated. (ii) It is natural in (psycho)linguistics to talk of learners hypothesizing grammars rather than languages; the latter is more common in computational linguistics. Reference to languages is often more convenient in discussion of SP, where talk of grammars makes for cumbersome locutions (a grammar which generates/licenses a language which is a subset of the language generated/licensed by another grammar).

language learning. The conclusion must be that children have some means of either avoiding or curing superset errors. Many learning models invoke SP for this purpose: it is intended to prevent superset errors by prohibiting LM from hypothesizing a more inclusive language than is warranted by the evidence.

The precise definition of SP will be a focus of section 3.1. An informal version which will serve well enough until then is that LM must never hypothesize a language which is a proper superset of another language that is equally compatible with the available data. As Clark (1992: 101) puts it: 'Given that the learner has no reliable access to negative evidence, it appears that the learner must guess the smallest possible language compatible with the input at each step of the learning procedure'. As is standardly noted, the effect of this is to guarantee that if a wrong language is hypothesized, that language is either a subset of the target or intersects with the target, and in either case there will be at least one target language sentence that the learner could encounter which is not licensed by the wrong grammar and so could serve as a trigger for change to an alternative hypothesis.

Doubts have been raised as to whether children do apply SP. There are periodic reports in the literature of children overgenerating and then successfully retreating to a less inclusive language (e.g., Déprez & Pierce 1993). (Other cases may occur unnoticed if the retreat happens prior to production of relevant sentence types in the child's speech.) Merely the occurrence of overgeneration errors is not decisive on this point. Even a learner designed to respect SP might overgenerate due to misparsing the input, or noisy input, or resource limitations, etc. However, if it is true that overgenerating children later eliminate their errors, that implies that LM contains some procedure(s) *other* than SP for arriving at subset languages; and if these retreat procedures exist, then SP might be unnecessary. Possible mechanisms of retreat include periodic sampling of less inclusive languages (contra the error-driven working assumption A7 above; see Briscoe 2000, Yang 2002); statistical measures of the frequency of constructions encountered; the past success of entertained hypotheses; and/or 'indirect negative evidence' as is created by preemption mechanisms such as Randall's (1992) 'catapults' or a 'uniqueness principle' that favors languages in which each proposition is expressible by only one surface form. However, working assumption A11 limits discussion in this paper to learning mechanisms *without* any reliable means of retreat from superset hypotheses. This leaves *avoidance* of superset hypotheses as the only strategy, and it is SP that is standardly claimed to bear this responsibility.

SP is of no interest to psycholinguistics unless there exist *subset–superset* (henceforth: *s–s*) relations among learnable human (UG-compatible) languages. It seems clear that there are. It is surely true, for example, that a natural language could exist which has no subjunctive but which is in all other respects just like a current variety of English that does have

subjunctives. Similarly, it seems probable that UG would tolerate a natural language exactly like some current variety of Italian but with the 'split DP scrambling' of Latin (adjective separated from the noun it modifies). A documented example of an s–s relation is the expansion of the *'s* genitive since late Middle English, which apparently was not accompanied by any (relevant) contraction of the language (e.g., no loss of the *of*-genitive), so the outcome was a superset language. Any real-life historical change consisting of loss or addition to a language is sufficient proof that one natural language can be a proper subset of another, and hence that human learners do sometimes face subset/superset choices.[4]

1.4 *Unresolved issues*

Although SP is essential for successful acquisition on the assumptions above, it faces problems of definition (which grammar choices *should* it favor?) and problems of implementation (how could a human learner apply it?). We will consider implementation problems first, in section 2. During that discussion we will presuppose the familiar interpretation of SP, as insisting on the smallest possible language compatible with the input (see Clark's characterization above), though we will subsequently (in section 3) challenge this as insufficiently precise.

We will address the following questions in turn. Can LM know when it is faced with an s–s choice? How economically can information about s–s relations between languages be mentally represented? Can this information be efficiently accessed and employed? (The answer to all of these will be 'possibly yes'.) When SP is applied, does it indeed facilitate acquisition? (Answer: yes in some respects but no in others.) Can undesirable consequences of SP be brought under control? (The answer is still unknown.) To the best of our knowledge there is no extant learning model which succeeds in implementing SP while remaining within reasonable human resource limits. Either SP problems are glossed over (as in our own prior work), or the solutions proposed would not be realistic within a psychological model (e.g., Chater & Vitányi 2005, submitted; see also Kapur 1994). The most intensive discussions of the issues to date remain those of Berwick (1985) and Manzini & Wexler (1987; also Wexler & Manzini 1987) almost 20 years ago. But as we note below, that work did not address the full range of s–s relations that natural languages may exhibit, nor did it note the unhappy conflict between SP and incremental learning that we discuss below.

---

[4] It does not follow that there are subset/superset parameters. The relations between linguistic constructions and parameters can be very indirect. See section 2.2 on s–s parameters. Also, we assume here that s–s relations with respect to syntax are not dependent on particular lexical content, though this is clearly an oversimplification if some parametric choices are tied to overt morphological realization of functional heads.

In order to facilitate the drawing of connections between formal learnability results and linguistic theory, we will consider the issues both in terms of languages (sets of sentences) and in terms of grammars, specifically grammars consisting of UG principles and values for UG-defined parameters. Except where otherwise specified, the principles can be assumed to be the Government-Binding principles of the original principles-and-parameters framework of Chomsky (1981), though these have been reconceived since. However, the points we make here are translatable into the terms of non-transformational and/or non-parametric syntactic theories, including versions of phrase structure grammar, categorial grammar, construction grammar, and others.[5] Though details will vary, SP issues arise in one guise or another for most theories of language and language acquisition. It is to be hoped that a strong theory of possible grammars will assist in solving SP problems, though it is conceivable that it will exacerbate them.

2. Implementation of SP

In order to make decisions which respect SP, LM must be able to recognize when a subset/superset choice presents itself. That LM has access to this information is often assumed without discussion, but it is far from obvious how it can be achieved. At worst, it could require LM to know (innately or by computation) the full range of candidate grammars compatible with its data, and all the s–s relations among them. Is this feasible? There seem to be three broad alternatives. (i) LM might directly compare the sets of sentences that constitute the candidate languages. Or (ii) LM might be innately equipped with a specification of all language pairs that stand in s–s relations. Or (iii) LM might be able to compare the *grammars* of the candidate languages, and choose between them on the basis of some general formal criterion. We consider these in turn.

(i) Could LM directly compare the sets of sentences that constitute the competing languages? This type of *extensional* comparison has the advantage of being a fully general method, equally applicable to all languages, but has the disadvantage of requiring on-line calculations that are implausibly complex.[6] Each learning event (each encounter with an input sentence) may

[5] Optimality Theory assumes that grammar acquisition consists of establishing the correct priority-ordering of innate constraints. This is a very different perspective and we cannot include it in the present discussion. See Prince & Tesar (2004) and references there for discussion of SP in a learning model for OT grammars. Keller & Asudeh (2002) discuss OT syntax acquisition.

[6] The task may be worse than complex. Joshi (1994: 510f.) notes that 'if the grammars are context-free or more powerful than context-free then the problem of checking subset relationships is, in general, undecidable'. Later he observes that 'the problem of checking subset relationships for the tree sets of context-free grammars is decidable. Therefore, if we assume that the learner is dealing with context-free grammars, then the SP can be

entail many such language comparisons. Before adopting a grammar hypothesis, LM would in the worst case have to check *every* subset of the language generated by that grammar, to determine whether it was compatible with the available data. But it seems unlikely that children engage in such procedures each time a novel input sentence calls for a change of hypothesis.[7]

(ii) Are learners innately equipped with a specification of all language pairs that stand in s–s relations? What form would such a specification take? It might be a brute list of relevant pairs, or be encoded as a table in which the information is more easily accessed. A possible implementation that we will have occasion to return to later in this paper is as part of an innate *enumeration*, a listing of all possible languages/grammars ordered in a way that respects SP: all subset languages appear earlier in the ordering than their supersets. A learner could obey SP by considering languages in this order and adopting at each step the next one that fits the data. This has been a familiar idea in computational learning theory since Gold (1967). The innate ordering might or might not be a *maturational* sequence, such that some languages cannot even be contemplated by the learner until a certain level of biological maturity has been reached (see Bertolo 1995a, b; Wexler 1999).

Regardless of the format in which the information is stored, any sort of innate specification of s–s relations might appear to presuppose some sort of evolutionary miracle which ensured that no s–s instances were omitted or inaccurately represented in the biologically encoded information in the infant brain. Fortunately, this is one aspect of SP-implementation that is *not* as troubling as it may seem. If any subset language were not accurately represented as such, it would not be given priority over its supersets by LM, so it would not reliably be acquired and would not survive in a language community. We as linguists would know nothing of it. The acquirable languages, on present assumptions, include only those for which accurate s–s information *is* laid down in infant brains (whether by accident, natural selection, or otherwise).

A more apt complaint against proposal (ii) has to do with scale. Even if the number of possible natural languages is finite, and even in theoretical frameworks designed to be maximally restrictive, the estimated number of possible human languages is very large.[8] The number of s–s relations among

---

instantiated, in principle'. See also Berwick (1985, chapter 5) on decidability issues related to SP.

[7] The pros and cons of (i) have been debated in the literature. For example, White (1989: 148) acknowledges this computation load problem but nevertheless favors extensional computation of s–s relations on other grounds, specifically because of its explanatory value in modeling second language (L2) acquisition. It 'suggests that learning principles and UG may be in different "modules"', which 'allows for the possibility that UG is still available to L2 learners but that the Subset Principle is not'.

[8] The number of possible languages is easiest to estimate by reference to parameters of language variation (see discussion of P&P theory below). Thirty independent binary

these languages might be so great that any pair-by-pair specification would exceed plausible psychological bounds. However, if there were some pattern to the s–s relations, the information could be mentally represented more economically. This is an interesting prospect that we return to below.

(iii) Could LM examine the competing *grammars*, i.e., make *intensional* rather than extensional comparisons?[9] Is there a formal property of grammars that would reveal which stand in s–s relations? This is an attractive possibility which holds promise of eliminating the workload excesses of alternative (i), while minimizing the extent of innate programming needed for alternative (ii). It amounts to the postulation of an *evaluation measure*, as proposed by Chomsky (1965). The theoretical challenge is to identify the particular evaluation measure that human learners apply. Chomsky proposed a simplicity metric over grammars. This would be explanatory, since it presupposes that human learners are 'least-effort' mechanisms, doing no more work than the input requires of them.[10] Chomsky emphasized that what counts as a simple grammar for LM is not *a priori* evident to linguists but depends on the format in which grammars are mentally represented. Discovering this representation format was to be a major goal for linguistic research. As matters turned out, no such theory was ever developed. This was not for lack of interest, but for principled reasons that are now apparent in hindsight and are discussed below. The frustrations of implementing this approach to learning were among the motivations for turning to the theory of principles and parameters (P&P; Chomsky 1981), which sidesteps the problems that arose in the attempt to formulate a plausible evaluation metric. In the next sections we first sketch briefly why the evaluation measure approach failed, and then examine why P&P theory was not similarly afflicted. Later we show that significant problems still remain.

---

parameters yield more than a billion languages; every 10 additional parameters multiplies the number of languages by approximately 1,000. An aim of linguistic research is to trim the number of languages by uncovering UG constraints that tie several kinds of superficial language variation to the same underlying parameter. Work along these lines has had some successes, yet the number of parameters postulated has tended to increase rather than decrease as research proceeds. The scale of the language domain is relevant to how LM might be designed. Though no specific cut-off can be established *a priori* for biological feasibility, an arbitrary innate listing of grammars must become less plausible as its size expands.

[9] An intensional (I-language) definition of SP in terms of grammatical derivations is offered by Wexler (1993). However, it does not lend itself to being operationalized by LM in terms of comparing the competing grammars. Dell (1981) implies that LM could obey grammar-based (intensional) principles which follow from SP (such as: favor obligatory rules over optional ones; maximize bleeding relations between rules and minimize feeding relations), but he does not claim that these exhaust the content of SP. See Berwick (1985) for a similar approach in terms of the acquisition of syntactic parsing procedures.

[10] Chater & Vitányi (2005, submitted) assign a very different role to a simplicity metric in their model of syntax acquisition, but they do not present it as psychologically feasible.

2.1 *Simplicity and generality*

It was hoped that SP would fall out from a general simplicity-based evaluation measure over grammars, giving subset hypotheses priority over superset hypotheses. But the simplicity metric approach was undermined by the fact that in almost any natural format for representing grammars, there is a positive correlation between simplicity and generality (see discussion in Fodor & Crain 1987). Simpler representations are more underspecified and hence are more general in their application (this point is emphasized in H&R's paper though they come to a different conclusion from the one we will draw). This is by design; it has been a central tenet of linguistic theory not just for descriptive elegance but also under a psychological interpretation. In an early treatise, Halle (1962) argued on these grounds that phonological features are psychologically real and phonemes are not. His premise was that the representation system for grammars must explain why natural languages favor broad rules (e.g., /a/ becomes /æ/ before any front vowel) over narrower ones (/a/ becomes /æ/ before /i/). In feature notation, but not in phoneme notation, the broader rule is simpler, so it would be favored by learners employing a simplicity metric. Note that a positive correlation between simplicity and generality is regarded here as advantageous: it explains *why* learners generalize (a paramount concern at that time, in the battle against behaviorism). Similar reasoning would apply to syntax. However, our goal here is exactly the opposite: it is to explain what prevents learners from *over*generalizing. For this purpose, Halle's conclusion would have to be turned on its head: a phoneme-based notation would have to be regarded as superior because it penalizes overgeneralization.[11] Thus, there is an inherent opposition between these two aims of learnability theory: to encourage but also to limit learners' generalizations. It is conceivable that both could be satisfied simultaneously by some grammar format such that simplicity entices learners to generalize up to some particular point and no further, but no such notational system is known.[12]

---

[11] It might seem that overgeneration due to the simplicity/generality correlation could be circumvented by giving up rules and adopting constraint-based grammars. A constraint is a negative grammar statement. The simpler and more broadly it is stated, the more sentences it excludes, resulting in a *smaller* language in accord with SP. But this is no solution. Negative grammar statements cannot be learned without negative evidence, so all constraints must be innate. Therefore, the language-specific facts that learners acquire must either be formulated in positive terms, as rules or lexical entries that interact with the constraints, or else must be captured by weakening the initial-state constraints. Though the latter is worth investigating, it may fall victim to the same risks of overgeneration as noted above; see Fodor (1992) for discussion.

[12] One attempt to prevent least-effort learning from overgeneralizing assumes innate feature defaults, which apply to simple underspecified rules *in the grammar* and fill in specific properties *as the rules are applied*. Broad generalizations would require extra specifications in the grammar rules to override the specific defaults (Fodor 1992). However, while reining in the generalizing power of the notation, this errs in the other direction: it makes the

In the introduction to the classic collection of papers in Roeper & Williams (1987), Williams (1987) cites this failure of the simplicity metric as a reason for turning away from rule-based grammars, towards a parametric framework. But he emphasizes that parameter theory is under the same obligation to explain how learners order their hypotheses in such a way that they conform to SP. In the next section we consider ways in which this might be achieved.

2.2 *Does parameter theory help?*

Parameter theory broke out of the simplicity/generality tangle by shifting to an entirely different format for grammars, for which the notion of representational simplicity is irrelevant. The syntactic component of a particular P&P grammar consists exclusively of a collection of parameter values. Chomsky (1981, 1986) portrayed parameters as like switches. Since one setting of a switch is neither simpler nor more complex than the other, a simplicity metric over grammars would be inapplicable. However, if SP is to be satisfied, some prioritization of grammar hypotheses is still needed. This might be of type (ii) above: an innate mental inscription of all s–s pairs. Or it might be of type (iii): a general metric over P&P grammars that orders subset hypotheses before superset hypotheses, so that it would take less effort for LM to adopt the former. Let us consider what this metric would be like.

An appealing proposal is that all it takes to capture s–s relations between languages is to assign a default value to each parameter. This was contemplated in the foundational discussions by Manzini & Wexler (1987) and Wexler & Manzini (1987) and versions of it have been adopted by other scholars. (For exposition and commentary, see for example Atkinson 2001.) The points we make here diverge in some details from Manzini & Wexler's, in ways we will explain as we proceed, but our discussion in this section and below clearly owes much to theirs. Following their lead, we consider two propositions here: that some (if not all)[13] parameters have a default value (which we designate `0`) which licenses languages that are subsets of the languages resulting from the marked value (designated `1`);[14] and that

---

eventual (adult) grammar unduly complex. A remedy sometimes proposed is 'restructuring' of the learner's grammar at some point, to reduce complex conservative formulations to simple ones once the danger of mislearning is over. But a theory of restructuring also has its problems, e.g., how can a learner know when it is safe to undertake?

[13] Manzini & Wexler argued that parameters can be multi-valued, but we will assume only binary-valued parameters here. Also, Manzini & Wexler explored the proposition that the values of *every* parameter define s–s relations; this is their *Subset Condition*. As noted below, we assume, contra the Subset Condition, that there are some non-s–s parameters in the natural language domain.

[14] Nothing ensures that the default value dictated by SP is the value that linguists would regard as unmarked on grounds of linguistic 'naturalness', frequency across languages, etc. As one example: for topicalization, SP would force obligatory topicalization to be the

there are no s–s relations in the natural language domain except those which result from this. Let us call the conjunction of these two propositions the *Simple Defaults Model*. The s–s information coded in these parameter value assignments would not be taxing for a learner to apply on-line. LM would start in the maximal default state with all s–s parameters set to 0, and would reset a parameter to 1 only when the input requires it.[15]

For this strategy to be failsafe, the s–s relation between the 0 value and the 1 value of a parameter must be fully reliable, i.e., *constant across all combinations of values of the other parameters*, even if they are varying at the same time. The 0 value must *never* give rise to a superset of a language licensed by the 1 value. This is a variant of Manzini & Wexler's important *Independence Principle* (Manzini & Wexler 1987: 435). When we refer to *independence* in the discussion below, it is this version we intend unless specified otherwise. It is stricter than Manzini & Wexler's in one important respect: it imposes consistency of the s–s relation between the values of a parameter even if more than one parameter is reset at a time (which Manzini & Wexler's learning model apparently did not permit). In another respect, our independence principle is weaker than Manzini & Wexler's, since it does not require that if the 0 value of a parameter ever yields a proper subset of its 1 value, it must do so in all other parametric contexts. Rather, it requires only that a subset–superset relation between the values of a parameter is never *reversed* to a superset-subset relation in the company of any other values of the other parameters.[16]

---

default, and optional topicalization the marked value, though some linguists would make the opposite judgment. Potential mismatches between learning defaults and linguistic markedness criteria are another problem raised by SP but we must set it aside here.

[15] It is very important to note that this does *not* mean that LM is always justified in setting a parameter from 0 to 1 on encountering a sentence that is incompatible with the 0 value and compatible with the 1 value. Only an *unambiguous* trigger (as Manzini & Wexler seem to have been assuming) for the 1 value of that parameter would justify such a move and avoid risk of a superset error. But unambiguous triggers are not common; many natural language sentences are ambiguous with respect to which parameter values could license them (see Gibson & Wexler 1994, Fodor 2001, Fodor & Sakas 2004). Ambiguous triggers do not necessarily result in superset errors, but they do call for special caution, as we discuss in section 2.3.

[16] This is necessary if s–s parameters are to permit the existence of properly intersecting languages. Consider two s–s parameters, P1 and P2. The language with parameter values 01 (i.e., parameter P1 set to 0 and parameter P2 set to 1) is by definition a proper superset of language 00 and a proper subset of language 11; the same is true of language 10. Now what relation holds between 01 and 10? The independence principle, we assume, does not permit 10 to be a proper subset of 01, because that would be a reversal of the s–s relation between the values of P1, and LM would be unable to rely on preferring 0 values as a way of avoiding superset errors. For the comparable reason, 01 must not be a proper subset of 10. If these languages weren't permitted to intersect, either they could not both exist, or they would have to be disjoint. Our independence principle does permit them to intersect, since that does not threaten the strategy of preferring 0 values. (See Bertolo 1995b, for some results concerning the degree of dependence among parameters.)

The natural language domain apparently also contains parameters that do not engender s–s relations among languages, such as the null subject parameter as characterized by Hyams (1986) with triggers for both of its values; or headedness parameters (e.g., verb-initial versus verb-final VP) in theories that don't assume a universal base; or parameters for obligatory movement.[17] Languages that differ with respect to such a parameter are either disjoint or intersecting. As observed below (section 2.3), it may be helpful for these non-s–s parameters to be arbitrarily assigned a default value (i.e., a starting value). It is important, though, that their non-s–s status be known to LM, by some means or other, because resetting a non-s–s parameter should always take priority over resetting an s–s parameter, in order to avoid superset errors.

The Simple Defaults Model claims that the *only* s–s relations in the natural language domain are those which fall out from the supremely simple representation system of 0/1 values for each individual parameter. For example: if there were just five parameters, P1–P5, all of them s–s parameters, the language of grammar 01101 could have as its subsets only the languages 00101, 01001, 01100 and their respective subsets 00001, 00100, 01000, 00000. The Simple Defaults Model for representing s–s relations thus greatly limits the number of such relations that natural languages could exhibit. For instance, it entails among other things that languages with the same number of non-default values could not be subsets or supersets of each other; e.g., 00001 would not be a subset of 00100, or vice versa.

If the Simple Defaults Model is to solve the SP-implementation problem, the limited set of s–s relations that it is capable of encoding must exhaust the s–s relations that exist in the natural language domain. If the model could capture all *and only* the actually occurring s–s relations, that would be even more impressive, a considerable tribute to the explanatory value of the P&P theory. A glance at some examples suggests, however, that the Simple Defaults Model may be too limiting. It renders unlearnable a language domain such as in figure 1, where the four languages defined by two parameters are nested one within another. (Note: Though we show only the effects of these two parameters, they should be thought of as functioning within a larger domain.) If LM begins with grammar 00, and then encounters input *s*, it must clearly set parameter P1, not parameter P2, forward

[17] Such parameters are ruled out by Manzini & Wexler's *Subset Condition*. See Atkinson (1992) for extensive discussion of the Subset Condition. It has generally been embraced with less enthusiasm than the Subset Principle. For instance, MacLaughlin (1995: 187) maintains that 'UG parameters (to date) simply do not seem to meet the Subset Condition'. However, MacLaughlin then concludes that 'the Subset Principle has not been shown to be necessary to guarantee successful acquisition of UG parameters.' As will be discussed below, even if it were true that no individual parameter creates an s–s relation, that would not make it unnecessary to solve the problems SP raises, since s–s relations can arise in other ways (e.g., through lexical entries, or combinations of parameter settings).

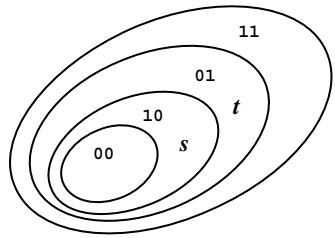


Figure 1
A possible s–s relation not compatible with independence: the 1 value of P1 yields a superset relative to its 0 value in one case (10 is a superset of 00) but a subset relative to its 0 value in another (10 is a subset of 01)

to its marked value, since the latter would result in a superset error if the target were 10. Later on, if LM encounters input *t*, it can set P2 to its marked value. (At that point, P1 must be set back to 0 in order to avoid a superset error if the target is 01. This aspect of the learning situation is the topic of section 3.) Thus, when LM has a choice of which parameter to set, P1 must be given priority over P2.

The Simple Defaults Model cannot establish the necessary priority relation *between* the two parameters. It requires only that for each parameter, LM must adopt the 0 value before its 1 value, so it offers LM a free choice of which parameter to set to 1 first when the input permits either. But in the figure 1 situation, a superset error cannot be avoided just by a strategy of favoring 0s over 1s. Hence, the Simple Defaults Model cannot provide LM with the information necessary for obeying SP if any part of the natural language domain is configured as in figure 1. Prioritization of one parameter relative to another would be needed also. Let us call a system that permits this (i.e., ordering of parameters as well as ordering of values within each parameter) an *Ordered Defaults Model*. This is not a priori implausible: in a P&P framework UG must in any case specify the parameters, so it might specify them in a fixed sequence. A natural language grammar would then consist of a *vector* of parameter values, in which the ordering is significant. LM would follow the strategy of hypothesizing marked values only as directed by this parameter vector, adopting the first grammar in the ordering that is compatible with the input.

Whether or not ordering of parameters is needed to capture s–s relations in the natural language domain is for empirical and theoretical linguistic research to determine, but a best guess at present is that counterexamples with the character of figure 1 can arise. One way in which they could do so would be if P1 and P2 license optional constructions that are not surface (string) distinct from each other, but occur in a narrower versus broader set of contexts. We might imagine that the marked value of P1 controls optional Wh-movement (versus none) while the marked value of P2 controls optional scrambling of XP (including Wh-XP) into a comparable position (versus no

scrambling). Then, as in figure 1, SP-compliance would require that in response to a sentence with a fronted Wh-XP, LM should reset P1 rather than P2; a later encounter with a fronted non-Wh item could trigger resetting of P2. However, a learner guided solely by choice of `0` over `1` within each parameter would have no way to know it should give priority to Wh-movement. Of course, no particular linguistic example is definitive. Its status depends not just on the language facts but on the proper description of the facts. Such cases might be dismissed on the grounds that movement operations are never truly optional,[18] or that the landing sites for Wh-movement and scrambling differ and LM could always tell them apart. Other potential cases might prove to be eliminable by a re-parameterization of the language facts (cf. Frank & Kapur 1996) which reallocates the descriptive responsibilities among parameters in such a way that a situation as in figure 1 doesn't arise. Imagine now that the marked value of P1 in figure 1 allows local scrambling (versus no scrambling) and the marked value of P2 allows scrambling unconstrained by locality (versus no scrambling); then P1 would need to take priority over P2 as in figure 1. But no prioritization would need to be externally imposed if the situation were re-parameterized so that P1 controls scrambling versus none, and P2 offers a strong versus a weak locality constraint on such movements. Alternatively, if multi-valued parameters are permitted (as by Manzini & Wexler), then a single 3-valued parameter could cover the two degrees of scrambling as well as none, so it would suffice to prioritize the values within that one parameter, without parameter ordering.

It has been argued by Dresher & Kaye (1990) and Dresher (1999) that parameter ordering is essential for phonological acquisition. For syntax, parameter ordering has been proposed on psycholinguistic and computational grounds. It is said to account for empirical observations indicating that the sequence in which children set syntactic parameters does not always reflect the sequence in which the relevant input triggers become available to them (e.g., van Kampen 1997; see also Borer & Wexler 1987 and Felix 1992, on the 'triggering problem'). It has also been invoked to explain children's immunity to acquisition traps they might have been expected to fall into (see Roeper & de Villiers 1992). Also, ordered parameters have been included in computational models for various purposes. In the *Triggering Learning Algorithm* (*TLA*) of Gibson & Wexler (1994) parameter ordering is contemplated as a means of avoiding crippling local maxima in hill-climbing learning; see extensive discussion in Bertolo (1995a, b). Briscoe (1999, 2000) orders

[18] Optional rules are a more common source of s–s relations than obligatory rules. There has been recent interest in eliminating options within transformational derivations (Chomsky 1995 and since) for theoretical and computational reasons. It will be important to determine whether this has the effect of shrinking the class of UG-compatible languages in a way that significantly reduces the number of s–s relations in the domain.

parameters from more general to less general as part of a statistical retreat mechanism based on preemption.

If parameter ordering turns out to be unavoidable for natural language, that means we must give up parameter independence. In figure 1 it is clear that parameter P1 violates independence. When the value of P2 is held constant, P1 is well-behaved (`00` is a subset of `10`, and `01` is a subset of `11`), but a reversal occurs when P2 also varies in value between the two languages in question (language `10` is a subset of language `01`); this is why LM cannot avoid overgeneration just by favoring `0` values over `1` values. Thus, while the Simple Defaults Model assumes independence of parameters, the Ordered Defaults Model does not. All the s–s relations captured by the Simple Defaults Model in the five-parameter illustration above would still be captured by the Ordered Defaults Model, but others would be as well. For instance, language `00000`, with no marked values, would still take priority over all other languages as before, but now (if we adopt a left-to-right ordering convention pro tem) the language `10000` would take priority over `01000`, which would take priority over `00100`, and so forth; and all of these would take priority over `11000`, `10100`, etc. with two marked values each. Obviously this is just one of many possible schemes for reading off a sequence of grammar hypotheses from an ordered list of parameters and their values. For some version of this general approach to succeed, all that is required is that there be some well-defined (computable) function that is simple enough to be psychologically feasible and captures all s–s relations in the domain. More interesting ordering schemes could be imagined; for instance, reading the parameter vector as if counting in binary, or a hierarchical arrangement in which setting one (parent) parameter makes accessible a collection of related sub-parameters (Roeper & Weissenborn 1990; see also Briscoe 1999, 2000; Villavicencio 2001).

The Ordered Defaults Model for representing s–s relations encompasses many possibilities and so might prove to be overly powerful. In one respect, it clearly needs to be relaxed. For example, if it imposed a *total* ordering over all grammars, and if it were stipulated that *every* priority relation between grammars corresponds to a subset relation between the corresponding languages, then the language domain would be very odd. It would necessarily contain just one massive nesting of every language within the next largest one. To avoid imputing this to the domain of natural languages, it would have to be supposed instead that the ordering captures all genuine s–s relations in the domain but also includes cases where the ordering is irrelevant for purposes of SP. That is: the ordering would meet the condition that an earlier-ordered language is never a superset of a later-ordered language, but not the condition that an earlier-ordered language is a subset of every later-ordered language. For purposes of SP, then, the ordering of grammars is still only a partial ordering, though it is a more constrained partial ordering of

grammars than that defined by the Simple Defaults Model. (See further discussion in section 4.2.)

Our conclusion so far is that the need for knowledge of s–s relations may not, after all, be a serious obstacle to implementation of SP. In breaking away from rule-based grammars evaluated by a simplicity metric, P&P theory offered new ideas for how to prioritize grammar hypotheses so as to avoid premature adoption of superset languages. If either version outlined here succeeds in providing the information needed for SP-compliance, it would reveal something interesting about the s–s relations among natural languages. It would show that they are non-arbitrary under a parametric description: there must evidently be some linguistic regularity of a kind which permits the compression of this information into an order of (parameters and) parameter values. However, absent systematic research on this topic, we present it here as a possibility worth investigating, without taking a stand on whether it will ultimately prove adequate for natural language, or whether some other theory of grammars may offer a formalism that better captures the s–s information that LM needs. (See Wu 1994, for discussion in a Minimalist framework; and Kanazawa 1994, Briscoe 1999 and elsewhere, within a categorial grammar framework.)

In the next section, we examine a potential threat to *any* such s–s representation system. Trouble would arise if natural languages were found to exhibit complex or disorderly interactions among parameters, so that s–s relations could be captured only by some more intricate notation than parameter value vectors, or perhaps only by non-systematic enumeration of individual grammars.

2.3 *On parametric conspiracies*

The possibility of interactions among parameters – even parameters that are independent in the technical sense above – has been brought to attention in important work by Clark (1989, 1992). Clark (1992: 102) warned that:

> In brief, a learner could obey the Subset Condition [this is Clark's name for the Subset Principle, JDF/WGS] on the microscopic level (with respect to a single parameter) while violating it on the macroscopic level (due to shifting interactions between parameters).

We need to assess how damaging this 'shifting problem' could be. Would it preclude any orderly parameter-based solution to the problems of SP implementation such as we considered in the previous section?

Parametric interactions of the kind that Clark has highlighted (illustrated in figure 2) are not exotic; they are probably widespread in natural language. Their essential ingredients are merely: a target language that has a superset, and an input sentence that is ambiguous between the target and another language (which is not a superset of the target). The type of problem this can

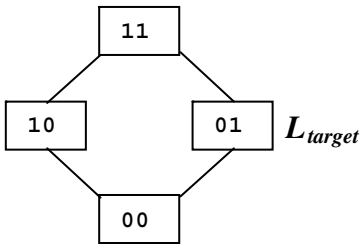


Figure 2
Based on Clark (1992): a language domain in which SP may be ineffective

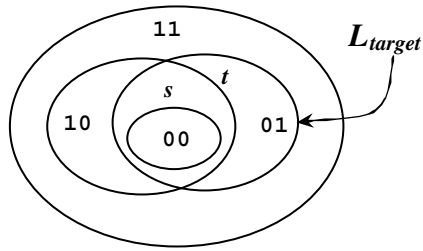


Figure 3
Language-inclusion representation of the language domain in figure 2

engender is illustrated by Clark (see also Nyberg 1992) by reference to parameters for structural case marking (SCM) and exceptional case marking (ECM). Figure 2 shows this example as diagrammed by Clark (1992); we have made only minor changes of notation. Nyberg (1992) presents the same example with linguistic content assigned to the two parameters: the marked value of parameter P1 licenses SCM, and the marked value of P2 licenses ECM. (If these assignments were reversed the same conclusions would follow; the relevant aspects of the example are symmetric.) The lines linking grammars in figure 2 indicate s–s relations: a lower language is a proper subset of a higher language that it is linked to in the diagram. For convenience, the same state of affairs is shown in figure 3 in terms of language inclusion. The target grammar is `01` (i.e., the target language has ECM, as in English, not SCM as in Irish). Note that both parameters here are s–s parameters.[19] Clark observes also that they satisfy Manzini & Wexler's Independence Principle, thus making it clear that that principle does not solve the shifting problem.

This example is commonly regarded as revealing the impossibility of applying SP online in incremental learning, since the parameters apparently conspire to entice LM into a superset hypothesis (grammar `11`, with both ECM and SCM), making the target `01` unattainable. If SP were operating properly, it should ensure that the learner never hypothesizes `11` in this situation. Clark's discussion raises a serious doubt as to whether SP can in fact block this error. However, on closer examination it becomes clear that such an example does *not* show the impossibility of applying SP on-line.

[19] Both P1 and P2 need to be s–s parameters in order to create Clark's problem concerning SP. By definition of the problem, the target (designated here as `01`) is a subset of the error hypothesis `11`. Thus P1 exhibits an s–s relation, and the independence principle that Clark was assuming (Manzini & Wexler's, not ours) then requires that `00` be a subset of `10`. Now consider P2. It needs to be an s–s parameter in order to create the challenge for SP. If it weren't, SP would direct LM to make a non-superset move from `00` to `01`, rather than a superset move from `00` to `10`. Since `01` is the target, learning would be completed without any problem arising.

Nor does it challenge the adequacy of a parametric format for the mental representation of s–s relations (so the guardedly optimistic conclusion of section 2.2 above still stands). What such examples do show is that LM must faithfully *respect* s–s relations. This sounds too obvious to be worth stating, but it has an unexpected twist to it. As we will explain below, there is no conspiracy here powerful enough to overrule SP when SP is conscientiously applied. But we will also show that conscientious application is not possible if LM *sets* individual parameters independently of each other – even for parameters that satisfy the independence principle.

We now track through the learner's sequence of hypotheses in the course of navigating the domain shown in figures 2/3, to see how a superset error could arise. The learner begins at language `00` with both parameters at their default values; this language has neither SCM nor ECM. An input *s* (see figure 3) is encountered in which a nonfinite complement clause has an overt subject. Since the subject must have received Case somehow, *s* invalidates grammar `00`, but it is ambiguous between `10` and `01` and also the superset `11`. What is the best course of action for LM? A truly cautious learner would decline to set either parameter on the basis of this ambiguous trigger; it would wait for an unambiguous trigger for one parameter or the other. However, this is not a practical option unless LM can *recognize* when a trigger is ambiguous, and on standard assumptions this is not feasible. In order to detect ambiguity, LM would have to check multiple grammars against sentence *s* (in the worst case it would have to check *every* grammar in the domain) to find out whether *s* is licensed by only one grammar or by more than one.[20] On the assumption that this exceeds any reasonable computational load, the discarding of ambiguous triggers is not a solution to the shifting problem.

Lacking ambiguity detection, an alternative strategy would be for LM to adopt *any* grammar that licenses *s* and satisfies SP. On the Simple Defaults Model it could freely adopt either `01` (the target, in this example) or `10` (non-target). On an Ordered Defaults Model, LM's choice would be determined: if P1 takes priority over P2, it would adopt `10`. In either case (i.e., with or without parameter ordering), the learner might at this point have both parameters set wrong, i.e., its current hypothesis might be `10` instead of `01`. (Note: If the parameter priorities were reversed, a comparable situation would arise when the target was `10`; thus, parameter-ordering offers no

[20] Recognition of unambiguous triggers would not require this labor if they are innately specified for learners in the form of uniquely distinctive syntactic category sequences (e.g., N Aux V PP). It would be of enormous benefit to acquisition theory if this were the case, but linguistically realistic examples are hard to come by. For discussion of trigger ambiguity and how LM might detect it, see the references in fn. 15 above and Fodor (1998b). In our previous work we have proposed a learning model (the *Structural Triggers Learner*, *STL*) which does have some ability to detect and discard ambiguous triggers. But this is not essential in order for SP to weather the shifting problem, as we will show.

general solution to the shifting problem.) At some later time this wrong grammar 10 would be disqualified by an input sentence, *t*, that cannot be licensed by SCM alone. This could be a sentence in which the subject of a nonfinite complement is an anaphor bound by the subject of the matrix clause. This can only be analyzed as due to government of the subject by the higher predicate, i.e., as ECM.[21] With grammar 10 thus eliminated, LM's choice at this point is between 01 with ECM only, and the superset 11 with SCM and ECM also. Since the shifting problem is not intended to rest on mere ignorance of s–s relations, it is to be assumed that LM knows that 11 is a superset of 01, in which case SP would guide LM to adopt 01 at this point since 01 is the smallest language compatible with LM's input. Since 01 is the target, learning is now complete. Note that it proceeded without any hitch: at no point was the learner tempted to hypothesize the superset language 11. What, then, is so troublesome about this example that makes it the focus of a special warning by Clark and Nyberg?

For Clark and Nyberg it *is* troublesome, because they make additional assumptions about LM which block the crucial intermediate move from grammar 10 to grammar 01. We wish to make it quite clear that these are assumptions Clark and Nyberg make about the operation of what they take to be standard incremental parameter-triggering models. After observing the problems such models face, Clark and Nyberg themselves advocated very different approaches to the setting of parameters, based on other assumptions (which we will not discuss). For standard incremental parameter setting, Clark assumed the Single-Value Constraint (SVC), which permits only one parameter to be reset in response to one input sentence (see Clark's (1992: 90) formal definition). Therefore, when P2 is set forward to 1, P1 cannot simultaneously be set back to 0 as SP requires it to be. Hence LM can shift from 10 only to the superset 11.[22] Nyberg assumed that standard parameter setting is deterministic in the sense of no backtracking, i.e., 'the learner … may not reset a parameter that it has already set' (p. 29). For a learner that starts with 0 values, this means it may never switch any

[21] In an extension of this example, Clark and Nyberg introduce the parameter for long-distance anaphora (LDA), noting that a subject anaphor does not in fact necessitate ECM but could be due instead to SCM plus a marked setting of the LDA parameter. This is of particular interest since it constitutes an ambiguity between a parameter in the case module and a parameter in the binding module, showing that modularization of the grammar cannot eliminate global considerations in grammar acquisition.

[22] Clark (1992: 97) suggests that a learner in this situation, not permitted to switch from 10 to 01, could nevertheless satisfy SP by backtracking to 00 (contra Nyberg's determinism assumption; see below); later it could move forward to 01 in response to some new input sentence. The danger then is that the new input might trigger the same unproductive series of steps, from 00 to 10 and back to 00 again, in 'a cycle of infinitely going back and forth in a series of hypotheses' (p. 98). An SVC learner would therefore need something to block endless revisiting of the same grammar. This would seem to require memory. See section 4.2 for discussion.

parameter from 1 to 0. Therefore, when 10 is found to be incorrect, LM has no choice but to move to the superset 11. Thus, the source of the problem is not any conspiracy between parameters, but the assumed determinism (unrevisability) of parameter setting. Deterministic parameter setting is too stringent a condition for natural language acquisition. It would be workable only if there were some unambiguous triggers and LM knew which they were. A parameter that has been set from 0 to 1 on the basis of a known-to-be-unambiguous trigger should indeed be locked into that value and not subjected to further changes. But in the more realistic case where triggers are often ambiguous between parameters, so that LM can't be certain whether a current value was accurately triggered or not, LM must be alert to the possibility that the current value might have been adopted in error and may need to be reset in response to subsequent input.

Clearly, then, the learning problem in figures 2/3 arises not because SP cannot adequately protect LM against superset errors, but because *SP is not reliably obeyed* by a learning algorithm in which some other constraint (SVC or determinism) outranks SP. When SP is flouted, overgeneration errors inevitably ensue. These examples therefore do not call for the abandonment of SP or the abandonment of a standard incremental parameter-triggering model. Or at least, they do not do so unless these additional constraints on LM can be shown to be necessary and to necessarily outrank SP; but it seems most unlikely that this is so. SVC has not proved very helpful for learning; see results by Berwick & Niyogi (1996) and Sakas (2003). So SVC could be abandoned. Or it could be weakened while retaining much of its original intent, which was to prevent sudden radical shifts of hypothesis. A milder form would be a constraint permitting only one parameter to be reset in direct response to an input sentence, but accompanied by as many other changes as are necessary to reconcile that one change with any non-negotiable constraints such as SP. Clark's example would then present no problem. In place of Nyberg's strict determinism, which prevents return to a parameter value previously hypothesized and then rejected, LM might obey a weaker but very natural form of determinism which forbids return to any *grammar* previously hypothesized (assuming the learner has memory for this, as Nyberg's model does; see also the discussion of memory in section 4.2). Then the shift from 10 to 01, needed for SP-compliance, would not be blocked.

There is, however, a highly significant point that is brought to light by Clark's example: LM cannot get away with *setting* a single parameter in isolation, as if it existed independently of the others. We have seen in this example that as LM sets one parameter forward in response to an input sentence, it may need to set others back in order to comply with SP. This is so even if the independence principle (Manzini & Wexler's principle or our variant of it) is satisfied by all the parameters in question. This is not a paradox. The fact is that there are two notions of independence here and they are not equivalent. One applies to grammars, and it limits dependencies

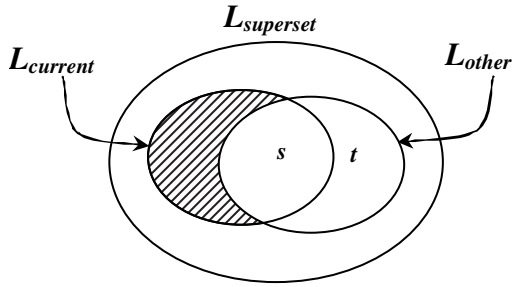


Figure 4
Sentences in the shaded area of $L_{current}$ must be given up when LM encounters input t which falsifies $L_{current}$

between the values of distinct parameters. The other applies to what learners know, and it regulates safe parameter-setting behavior. The two would fall together only if the learner's knowledge were unassailable, i.e., based exclusively on unambiguous input. If parameters are independent in the first sense (i.e., linguistically independent), then an unambiguous trigger for the marked value of a parameter does indeed – by definition – warrant the adoption and retention of that value regardless of the value of any other parameter in the grammar. But as we have noted, unambiguous triggers (known by LM to be unambiguous) are not the typical case. So in actual practice, LM has no secure grounds for holding on to marked values it once hypothesized, if a new parameter setting could cover the same facts. To insist on doing so would very clearly be a route toward overgeneration errors.

It emerges, then, that trigger ambiguity is the culprit which entails that parameter *setting* can't be independent even when *parameters* are. The challenge posed by Clark's example turns out to be how to model human learning in a way that is robust enough to obey SP when triggers are ambiguous – or when they even might be, for all that LM knows. In fact, it is a classic observation of learnability theory that LM needs to mop up after itself when revising a previous wrong guess that was based on ambiguous input, though this point seems to have become lost from view in the transition from rules to parameters. Figure 4 is a typical textbook illustration of this fact, from pre-P&P days.

In figure 4, $L_{current}$ is LM's present hypothesis, which it selected in response to the input s which is ambiguous between all three languages in this miniature domain ($L_{current}$, L_{other}, and $L_{superset}$). Next, LM discovers that $L_{current}$ is wrong, because it encounters input sentence t. LM must move from $L_{current}$ to either L_{other} or $L_{superset}$. The correct choice, required by SP, is clearly L_{other} (which properly intersects with $L_{current}$), not $L_{superset}$. Otherwise, an overgeneration error would occur if L_{other} were the target. But note that adopting L_{other} entails *giving up* some (perhaps many or most) of LM's presently hypothesized sentences in $L_{current}$. This is what we will refer to as

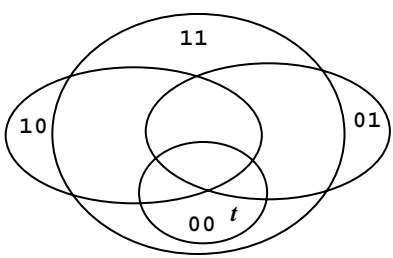
retrenchment. The moral of figure 4 thus clearly parallels the moral we drew from Clark's P&P example in figure 2, where we saw that the correct choice, required by SP, entailed giving up a marked parameter value at the same time as extending the language in another way. With or without parameters, then, it is clear that LM must engage in retrenchment. New hypotheses cannot simply be accumulated on top of old falsified hypotheses; the old ones must be discarded when LM makes advances in other directions.

Our discussion in this section has established two main points: (i) that SP can withstand the Clark challenge from 'shifting interactions between parameters'; (ii) that SP implementation requires systematic retrenchment in tandem with advancements. We'll return to point (ii) shortly. Point (i) is the good news for SP. Clark-type parameter interactions do not necessarily undermine the possibility of reliable SP-implementation, even for an incremental parameter setting learner. As long as LM has the requisite knowledge about s–s relations and *is not otherwise prevented from making use of it*, SP can be respected. Clark's examples do not represent true parametric 'conspiracies' which lie beyond the governance of SP in principle. Clark is right that SP violations could occur even if LM obeys SP with respect to a single *parameter* at a time, but this does not entail the more dramatic (and false) proposition that SP violations can occur over the course of learning even if LM obeys SP at each *learning step*, i.e., in its response to each input sentence.

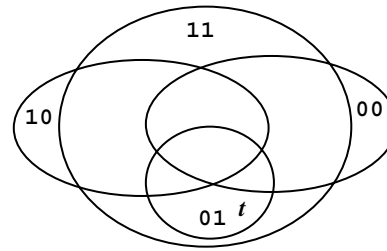
Are there language configurations which represent something more like a true parametric 'conspiracy'? It is certainly possible for an s–s relation to result from a combination of two (or more) parameters neither of which is an s–s parameter. For example, as diagrammed in figure 5a, `11` is a superset of `00` even though (unlike figure 3) neither P1 nor P2 alone yields a superset of `00`.

If such cases occur in the natural language domain, SP could get no grip on them as long as LM's knowledge of s–s relations is limited to an ordering of s–s parameters and their default (subset) values. To control such cases, it would appear to be necessary for LM to know about s–s relations beyond the reach of even the Ordered Defaults Model. But let us check whether this is so. Though this situation does present some potential problems, some potential solutions exist also.

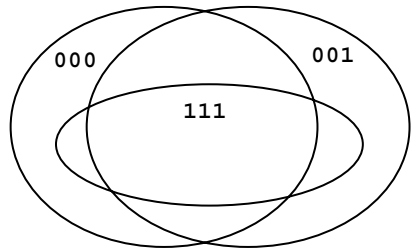
Suppose the target in figure 5a is `00`. Without SP to constrain the starting values for these non-s–s parameters, LM's starting grammar would not necessarily be `00`, so we may suppose it starts in `01`. Then in response to input *t* it might move to the intersecting language `11`, though this is a superset of the target from which it would never be able to retreat. For this learning problem there is a solution that is of some psycholinguistic interest: Arbitrary default (starting) values could be assigned to non-s–s parameters, thus *forcing* LM to start at `00` in figure 5a. However, some quite similar situations resist this solution. Since the assignment of defaults has no



a. L_{start} = 01, L_{target} = 00



b. L_{start} = 00, L_{target} = 01



c. L_{start} = 000, L_{target} = 111

Figure 5
Three language configurations that create potential learning problems

principled basis for non-s–s-parameters, the language configuration in figure 5b is just as legitimate as that in figure 5a. But now, even if LM is required to start at the default 00, it might fail to reach the target if the target is 01. It appears that on receiving *t*, LM could move freely from 00 either to the target 01 or to the superset error 11, both of which it intersects with. But safeguards exist for this case also. By our definition of independence (though not by Manzini & Wexler's), P1 counts as an s–s parameter in figure 5b. So SP would direct LM to give priority to resetting P2, and that would lead to the target 01, not to 11. Of course, this requires that LM knows that P1 is s–s and that P2 is not. Alternatively, the move from 00 to 11 could be blocked by the SVC (in the milder form we discussed above), which would forbid adoption of two marked settings (11) if one marked setting (01) would suffice to license the input. Other potential errors in the figure 5b situation could be blocked by enforcing retrenchment. For instance, if LM were to receive some other input sentence (not shown in 5b) on the basis of which it moved from 00 to the non-target 10, a retrenchment requirement would prevent it from subsequently moving from there to 11. Instead it would have to go via 01 (the target, by assumption), setting P1 back to 0 as it sets P2 to 1.

At present we know of no subset relations between languages that don't succumb to one or another of these natural constraints (defaults for non-s–s parameters; the weak SVC; the retrenchment constraint), but of course there may be some, lurking unnoticed so far in the natural language domain. And there are other interesting configurations of languages which, though they do not involve s–s relations, nevertheless resist learning on certain psychologically plausible assumptions. For instance, in figure 5c no language is a subset of any other, so SP offers no guidance. LM would fail to attain target language `111` if it were in the grip of an ambiguity-resolution principle (a principle determining its choice of hypothesis when more than one is available) that caused it to oscillate between hypotheses `000` and `001` in preference to `111` (e.g., a principle favoring a move to a more similar rather than a less similar grammar; note that the SVC is such a principle). Much comparative linguistic research will be needed before a reasonable assessment can be made of how perverse or cooperative the design of the natural language domain is from the point of view of psychologically feasible learning. If theoretically problematic relationships are observed among natural languages, learning theory will have to offer an account of how the human LM copes with them, and does so easily enough that the learning process does not lead to their elimination via language change. For the moment, in order to be able to move on to point (ii) above, we will simply assume, for lack of evidence to the contrary, that human learners face no choices which cannot be resolved in principle by SP in alliance with other plausible constraints.

Point (ii), the retrenchment requirement, is the springboard for the remainder of this paper. Clark's example shows that obedience to SP cannot normally be implemented one parameter at a time. Except in very unusual circumstances, LM must operate on the parameters conjointly: when it resets one, it must adjust the values of the others. This can have a significant impact on language acquisition theory. For instance, as noted above, any learning models that impose strong determinism or the unmodified Single Value Constraint are thereby eliminated from consideration, because they are incapable of complying with the retrenchment requirement imposed by SP. This includes some very familiar models, such as the TLA of Gibson & Wexler (1994). For models which survive this cut, the retrenchment requirement still exacts a cost since it undeniably complicates the parameter setting process. We will observe some of its more extreme manifestations in section 3.

3. How stringent is SP?

In section 2 we asked whether a psychologically plausible learning device could be equipped with knowledge of all subset–superset relations that hold between natural languages, and with a procedure for applying that

knowledge on-line as input sentences are processed and grammar hypotheses are revised. We found no principled obstacles to this, though many specific issues remain to be resolved. We now set aside all such issues concerning LM's ability to *apply* SP. To that end, we simply stipulate until further notice that LM has cost-free access to, and use of, all SP-relevant information about the languages in the domain. For instance, LM could freely retrieve and use information concerning the complete set of languages that contain its current input string; or it could determine the smallest language that is a superset of some other specified language; and so forth. Thus from now on, LM will be assumed to know everything it needs to know, except which language is the target, what input it will encounter, and what its past learning experiences have been (since by working assumption A5, LM has no memory for either past inputs or past hypotheses). This frees us to focus now on the question: Which language *should* LM hypothesize, in various circumstances?

3.1 *Defining SP*

Precisely what constraint does SP impose on the sequence of LM's hypotheses? On one hand, the answer is utterly familiar: subset languages take priority over supersets. On the other hand, SP has never, to our knowledge, been spelled out in such a way that its implications for *incremental* learning are clear.[23] Clarity is important because there is an ambiguity in the characterizations of SP to be found in the standard psychocomputational learnability literature. The ambiguity concerns the notion of relevant input data or triggering data; see the phrases boldfaced (by us; JDF/WGS) in the statements of SP below. Version (a) is due to Berwick (1985: 23); (b) is from Manzini & Wexler (1987: 414f.); (c) is the brief statement from Clark (1992: 101) that we cited earlier. The same ambiguity can be found in formal definitions of SP also. It is present, for example, in the variable *D* which is defined without further specification as denoting 'a finite set of data' in Manzini & Wexler's definition (1987: 433).

(a) **Berwick**: 'Briefly, the Subset Principle states that learning hypotheses are ordered in such a way that positive examples can disconfirm them. For many cases, the order will force the narrowest possible *language* to

[23] In section 4 we will pin learning problems on the lack of memory (working assumption A5) rather than on incrementality per se (working assumption A6). But the two are closely related, since memorylessness forces incremental learning (though not vice versa). In the meantime we will use the term 'incremental learning', which is more familiar in psycholinguistics than 'memoryless learning'. (In the literature of formal recursive-function theory, some kinds of incremental learning may be referred to as 'iterative learning'.) See Osherson et al. (1986), Lange & Grieser (2003) and references there for several theorems that establish classes of languages that are or are not in principle learnable by algorithms that are memory-restricted (though not strictly memoryless, since they typically presuppose an enumeration, which we classify below as a form of memory for past grammar hypotheses).

be hypothesized first, so that no alternative target language can be a subset of the hypothesized language. More precisely, no other target language **compatible with the triggering data that led to the new hypothesis language** can be a proper subset of that language.' (Clarification: For Berwick, 'target language' = any UG-permitted language; for us, 'target language' = the language the learner is exposed to.)

(b) **Manzini & Wexler**: 'The essence of the learning component is the Subset Principle, which orders parameter values according to the subset relations of the languages that the values generate. To be more precise, given two languages, one of which is a subset of the other, **if both are compatible with the input data**, the Subset Principle will state that the learning function must pick the smaller one.[24] The Subset Principle will then of course provide the solution to the subset problem.'

(c) **Clark**: 'Given that the learner has no reliable access to negative evidence, it appears that the learner must guess the smallest possible language **compatible with the input** at each step of the learning procedure.'

These characterizations of SP are adequate for learning systems with memory, which are able to scrutinize simultaneously all the inputs they have ever received. Then a phrase such as 'compatible with the input data' can be understood as referring to all data received since learning began. But that interpretation is impossible for an incremental learner without memory for past input. For such a learner the available data consists solely of the current input sentence. The smallest language compatible with that could be very small indeed – absurdly small. Though retrenchment is necessary, as argued above, here it seems too extreme. If this is the consequence of SP for incremental learning, it would amount to a reductio ad absurdum. It would have to be concluded either that SP does not apply to natural language acquisition, or that natural language acquisition is not incremental in this sense. In this section we assess how severe this problem is, before seeking solutions.

Is there any leeway in the definition of SP that could render it more friendly to a learner without memory? In (d) we give a formulation of SP which explicitly makes reference to how much of its past experience the learner has access to. While retaining the general spirit of (a), (b) and (c) above, this brings the memory factor out into the open so that we can experiment with it. As different specific memory capacities are assumed for LM, they will be entered into (d) and SP's consequences will vary accordingly.

---

[24] This definition *requires* LM to pick the smaller of the two languages in question, failing to allow for the possibility of LM's choosing some other language entirely, such as one that intersects with these. Berwick's definition, by contrast, requires only that LM not pick the superset language.

(d) When LM's current language is incompatible with a new input sentence *i*, LM should hypothesize a UG-compatible language which is a smallest superset of *i and all prior input sentences retained in its memory*, excluding any language *recorded in memory as having been disconfirmed by prior input*.

Note that we have coined here the technical concept *a smallest superset*, or more generally *a smallest language that meets criteria C*, which will be useful in the discussion below. Every language that meets criteria *C* and has no proper subset that meets criteria *C* is a smallest language that meets *C*. All smallest languages that meet the same criteria stand on equal footing from the perspective of SP. They do not stand in s–s relations to each other, so differences between them, such as their size, are of no concern to SP. Since SP deals only with s–s relations, it cannot favor one such language over another even if one is very large and the other is very small.

For a psychological agent, the formulation of SP in (d) is optimal, in the sense that a learner abiding by it will avoid both overgeneration and excessive retrenchment to the best of its ability to do so. It will give up just enough of its current hypothesized language as is necessary for safety, in view of its ignorance about aspects of the language which it has either not encountered or cannot now recall. For an LM with no memory limitations at all, (d) ensures that its selection of a grammar hypothesis will be based on all input received to date, and will exclude all grammars it has previously hypothesized to date (since, under the 'error-driven' working assumption A7 of section 1.2, and excluding issues of noisy input, all previously hypothesized grammars have been falsified). At the other extreme, if LM has only the very limited memory resources specified by working assumption A5, then (d) entails that its grammar hypothesis must be based solely on the current input sentence *i*, and can exclude only the current grammar which has just been falsified by *i*; thus, as stated in (d′), LM must shift to a smallest superset of $\{i\}$.[25]

(d′) SP for a memoryless learner: When LM's current language is incompatible with a new input sentence *i*, LM should hypothesize a UG-compatible language which is a smallest superset of $\{i\}$.

Note that (d′) calls for *maximum* retrenchment, eliminating as many sentences as possible in the course of adding the new input sentence *i* to the hypothesized language.

---

[25] Strictly speaking, LM could deduce that the target language contains not only *i* but also at least one sentence of its current language. This is because (except in the starting state) it could have arrived at its current language only in response to a sentence in the input, i.e., (setting noise aside) a sentence of the target language. However, LM has no way of knowing *which* sentence of its current language is in the target language, and it would be unwise for it to guess. We will therefore assume that SP enjoins LM to base its hypotheses on *i* alone.

How extensive would this retrenchment be in practice? In principle, the new language hypothesized under (d′) could be very small indeed (just one sentence) or it could be very large (e.g., infinite). All depends on the strength of the constraints imposed by UG. If the UG principles are powerful enough, they might require every natural language to contain certain sentences (or more plausibly, certain sentence *types*); if so, these sentences would not be eliminated during a retrenchment under SP. For example, perhaps all languages must contain active declarative single-clause (degree-0) sentences. In that case even the smallest language that LM is ever required to shift to would still contain active declarative degree-0 sentences.[26] Or UG might yoke two sentences (sentence types) together, so that any language that contains one must also contain the other. For example, UG might require that any language containing embedded interrogative clauses must also contain interrogative root clauses. Then interrogative root clauses would not be eliminated when LM encounters an embedded interrogative, though they might be eliminated in response to some other input sentence such as a root declarative or imperative. Or UG might require a language with relativization of oblique objects also to contain sentences with relativization of direct objects (cf. Keenan & Comrie 1977). Thus, the language the learner hypothesizes might not shrink drastically at every step. SP does not demand *full* retrenchment as long as LM has some other source of knowledge, such as UG, to supplement the meager information provided by its current input sentence.

Now we ask: how *frequent* would this retrenchment be? We have noted that SP in formulation (d′) for memoryless learners entails that the language that LM hypothesizes in response to an input *i* must contain only sentences (sentence types) that are *either* universal, *or* universally yoked to *i*. All other language facts that LM had previously acquired (correctly or incorrectly) must be given up when LM encounters *i* in the input stream. Now note that this must happen *every* time LM encounters *i*. If *i* is a frequent sentence in child directed speech (e.g., *It's bedtime*), it would trigger repeated retrenchment. For instance: presuming that not all natural languages have topicalization, we contemplate here a radical loss of topicalization from LM's hypothesized language *every* time LM encounters a sentence such as this one, which is neither topicalized nor yoked to topicalization by UG. This could happen even if LM were very advanced, on the verge of arriving at the complete target grammar. Of course LM could set about re-acquiring the phenomena it had lost, but it might not get much further the next time since a

[26] For brevity we are glossing over problems here. Even if LM is permitted by SP to retain its active declarative degree-0 sentences, their details (e.g., word order) in the currently hypothesized language may differ from those of the target. It is unclear how much can be retained in such a case (where UG says that such clauses must exist in every language but conservative learning warns that LM can't trust its current assumptions about their properties). Possibly LM keeps the construction but with default properties, but this solution would require further thought.

similar regressive pressure applies at every step. Cumulative acquisition is thus a major casualty of SP for memory-restricted learners.

Though we are unable at present to quantify exactly the extent of such repeated loss of past learning, it does seem likely to be a serious impediment to rapid acquisition. In order for LM to be absolutely sure that it will not overgenerate, it must err in the direction of undergeneration and so must settle for very limited forward progress. (Recall that this does *not* depend on any assumption of ignorance about s–s relations in the language domain.) It will be important for future research to consider whether this squares with empirical data on child language acquisition. Exactly how the strategies of a learning model could be expected to manifest themselves in children's behavior is a complex matter. Predictions are confounded by the host of other factors that enter into child development and performance. Nevertheless, it is worth pointing out that, as a first approximation, a model of incremental learning governed by (d′) portrays the incremental learner's route through the language domain as a case of snakes and ladders, with retreats as numerous as advances; and this seems unlikely to offer a good fit with the course of real language learning. In sections 3.2 and 3.3 below we will show that SP has even more serious consequences for incremental learners: without special assumptions, learning may fail entirely.

Because the consequences of SP under formulation (d)/(d′) are so dire for incremental learners, we may seek to edit it into something less punishing. What (d)/(d′) seems to overlook is the information available to LM in its currently hypothesized language – a language that it arrived at as a result of experience with the whole input sample to date. If that cache of information could be used to supplement the limited information contributed by the current input sentence, it might be possible to avoid excessive retrenchment. Unfortunately, as we now show, this is not possible.

We have been taking the memory-based formulation (d) to entail (d′) for an incremental learner. Now let us consider two alternative readings of (d) which are more respectful of LM's prior learning experience based on a larger sample of the target language. We will not, in the end, be able to endorse these, but since the stakes are high it is important to at least consider them.

(e) When LM's current language is incompatible with a new input sentence *i*, LM should hypothesize a UG-compatible language which is a smallest superset of its current language and *i*.

(f) When LM's current language is incompatible with a new input sentence *i*, LM should hypothesize a UG-compatible language which includes *i* and as many of the sentences of its current language as are compatible with *i*.

Interpretation (e) instructs LM always to expand its hypothesized language in order to accommodate new input. But we observed in section 2.3 above

that language contraction is essential on some occasions, as a means of recovering from a wrong hypothesis. This is the case not just for incremental learners but for *any* learner susceptible to hypothesizing non-target sentences – which is essentially all resource-limited learners faced with ambiguous input not recognizable as ambiguous. Sentences projected on the basis of now-disconfirmed grammars must be eliminated as the learner moves on to new hypotheses. Hence (e) is flagrantly at odds with the work that SP is supposed to do.

(f) is an intermediate proposal. It is safer than (e) since it acknowledges that LM must be prepared to give up some part of its currently hypothesized language as it absorbs a new input sentence. It is also less extreme than (d′), since it tries to *minimize* the loss of previously acquired sentences. This, however, is impossible. Two scenarios fall under (f). In one, all sentences in LM's current language are compatible with *i* (i.e., they and *i* co-occur in at least one UG-compatible language). Then (f) falls together with interpretation (e) and is untenable for the same reason. Suppose instead that some of LM's current sentences are incompatible with *i* (for some reason of UG which is not of concern here). If *i* is added to the language, those other sentences will automatically (via UG) be eliminated. Is the resulting language a safe hypothesis for LM? It may not be, if the shedding of old sentences does not go far enough.

Consider all and only the languages that contain *i*. In a dense domain there may be many of these. Some may be subsets of others. For all that LM knows, the target language could be a subset of any one or more of them. To be sure of avoiding overgeneration, LM's only safe move is to hypothesize a *smallest* such language. (See above: this is a language that contains *i* and has no subsets that do so.) If there is just one smallest language containing *i*, LM must choose it. If there is more than one, each perhaps at the heart of a different collection of nested superset languages, LM may choose freely among them. As noted above, from the point of view of safe learning, any such 'smallest language' is as acceptable as any other. Either way, though, the conclusion is clear. In the process of changing grammars to accommodate a new input *i*, LM must give up as many sentences as is necessary to arrive at a smallest language. As anticipated above, it must give up every sentence whose presence UG does not insist on in a language containing *i*. (f) does not require this and is thus too weak to do the work that SP is intended to do.

The conclusion must be that (d) does amount to (d′) for an incremental learner; there is no escaping this very strict construal. From this point on, we will employ (d) as our statement of the Subset Principle for a learner working under practical psychological limitations, and we will refer to it as *SP*(*d*). The fact that SP(d) entails (d′) for incremental learning is disturbing. The constant retrenchments it demands, and the learning failures that can result (section 3.2 below), are rarely if ever mentioned in the literature, but they are an important and troubling consequence of rigorous application of SP.

Informally: a learner with no memory for past learning events is like a newborn at every step of learning. It has no accumulated knowledge, so each new sentence might as well be the first sentence it has ever heard. Systematic progress toward the target is therefore impossible. Instead, the learner will shift from a smallest language containing its first input sentence to a (probably different) smallest language containing its second input sentence, and so on indefinitely. It could not consolidate its gains from previous experience. In short: Though a learner without adequate negative evidence has no choice but to apply SP, applying SP reduces forward progress to a crawl. Because this is so unwelcome a conclusion, we have taken trouble to show that more palatable alternative formulations of SP are not adequate, since they fail to accomplish SP's task of preventing superset errors. Some more radical solution is apparently necessary.

One radical solution would be to give up on SP entirely. SP can be rendered unnecessary by assuming, after all, that LM has access to direct negative evidence, or to some indirect form of negative evidence, or to retreat mechanisms, as we considered but set aside in section 1.3. Many researchers have declined this move because it would entail arguing that negative evidence (direct or indirect, in advance or in retreat) is available to correct *every* potential overgeneration error that *any* learner might ever be tempted to make. A linguistically more interesting path would be to try to control the extent of retrenchment by modularizing the grammar. The P&P framework would seem to be well-suited to this approach. If the parameters cluster into modules (e.g., case, theta, binding, bounding modules), it might be proposed that resetting one parameter need be accompanied only by retrenchment of other parameters in the same module. However, Clark (1992) showed that the case parameters for ECM and SCM (see section 2.3) and a binding parameter (for long-distance anaphora) compete in a way that could require retrenchment across module boundaries (see fn. 21). Perhaps instead there are not neatly defined modules but skeins of parameters, not obviously related to each other, which happen to compete only among themselves for licensing the input. LM could limit its retrenchments if UG were able to provide it with innate information about these relationships.

In a different vein, LM might be capable of determining on-line which parameters are relevant to the licensing of its current input sentence, and then it could restrict its retrenchment to those and not meddle with irrelevant parameters. There are at present no learning models capable of computing relevance relations between parameters and input sentences, because this presupposes 'decoding' of the input, i.e., that the learner can tell which grammars are compatible with any input sentence (cf. working assumption A9 above). Although decoding ability is often taken for granted in formal learning-theoretic discussions, how it could be psychologically implemented is rarely spelled out. Most current models (e.g., the TLA) rely instead on weaker tests of grammar/sentence compatibility which do not deliver a set of

relevant parameters but only a yes/no decision as to whether an antecedently selected grammar is able to parse the sentence. The *Structural Triggers model* (*STL*; see fn. 20) has some modest decoding ability, and so in principle might assess relevance. But exactly how to operationalize relevance relations is a thorny matter, and so far no fully satisfactory implementation exists. (For discussion see Sakas & Fodor 2001, Fodor & Sakas 2004.)

These are important possibilities for future research to pursue. In this paper, however, we will try out a different expedient. In section 4, rather than responding to (d′) by giving up SP or invoking modularity or relevance, we will experiment with ways of weakening its impact by providing the learning model with some memory. First, though, the full brunt of (d′) must be made clear. For a memoryless learner the penalty for a properly strict construal of SP is not merely slow and circuitous learning but can be total failure to converge on the target.

3.2 *Convergence failure due to SP*

In this section we note the possibility of a permanent failure to converge on the target grammar as a direct consequence of SP-compliance. Failure is not due to *over*shooting the target (as in superset errors), since the strategy of maximum retrenchment entailed by (d′) prevents that. Failure is due to constantly *under*shooting the target, as a result of the extreme caution that (d′) insists on. This recalls one prong of Gold's famous (1967) formal non-learnability proof: If the learner follows a conservative strategy in order to be able to acquire less inclusive languages, then the more inclusive languages become unlearnable. We show here that undershoot errors are not limited to abstract language domains such as Gold's but can also occur in the natural language domain, given working assumptions A1–A12.

3.2.1 *Undershoot errors due to retrenchment*

Under those assumptions, even a finite language domain defined by a small handful of parameters is not learnable unless every language in the domain contains a distinctive type of trigger, which we will call a *subset-free trigger*. The need for these special triggers follows directly from the fact that SP entails (d′) for incremental learning. Hearing an input sentence i, an incremental learner must posit a smallest superset of $\{i\}$. If the target is the *only* such language, learning is complete. If there is more than one candidate and the target is among them, then LM may guess between them and has at least a chance of attaining the target on each such occasion (as long as there is no bias against the target in its guessing strategy). Otherwise (i.e., if the target is not a smallest superset of $\{i\}$), LM could reach the target only on the basis of some sentence *other* than i – specifically, on the basis of a sentence of which the target language *is* a smallest superset. LM will never achieve the target if

there is no such sentence. It is evident that for a strictly incremental learner a *single* sentence must do the job, because LM has access to only one sentence at a time.

In short: a language L can be reliably acquired by an incremental learner only if L contains at least one sentence such that L is a smallest language containing that sentence. This sentence is a *subset-free trigger* for L. A subset-free trigger is not necessarily an unambiguous trigger, since it may occur in other languages which intersect with the target, and it will of course occur in all languages that are supersets of those languages or of the target. But a subset-free trigger for language L, even if not unique to L, is the *only* effective trigger for convergence on L by an incremental learner as characterized here which abides by SP. If such a trigger exists, and if the input is fair (working assumption A12) so that LM is guaranteed a chance of encountering it, then learning will succeed. Otherwise, it will not.[27]

Do natural languages have subset-free triggers? Independent of this learnability problem, there is no reason to believe that they do. Clearly, though, this is an empirical linguistic issue. Meeting the subset-free trigger condition would greatly restrict the class of learnable human languages. Restricting the language domain is a desirable theoretical goal, but a considerable amount of linguistic research would be needed to assess the validity of this restriction. We doubt that the answer lies here. Requiring subset-free triggers would exclude any language L such that for each of its sentences there is some other possible natural language that is a proper subset of L and contains that sentence. Among other things, this would rule out any natural language which is the union of two other possible natural languages. This condition appears to be too strong, since it would preclude a range of natural historical processes of addition and loss.

From a learning perspective, also, a subset-free trigger is a distinctly odd phenomenon, since it triggers the whole language at once from any prior state of learning. Before encountering it, LM is likely to make little sustained progress, due to repeated retrenchment with respect to all parameters (or rules or constructions) other than what the current input sentence exemplifies. However, it is of little import whether or not any interim progress in the direction of the target is made, because: (a) a subset-free trigger is still essential for reaching the target even if LM does happen to have crept close to the target along the way; and (b) once LM encounters the subset-free trigger, then (if that trigger is unambiguous or if LM is lucky enough to guess the right language on the basis of an ambiguous trigger) it will arrive at the

[27] These subset-free triggers bear close resemblance to the 'telltale subsets' of Angluin (1980). Angluin proved that language domains are learnable from positive data if and only if for every (target) language, L_i, in the domain, L_i contains a telltale set of sentences that is not contained by any other language in the domain that is a subset of L_i. Our subset-free triggers are analogous to a telltale consisting of a single sentence. (Note that there is no upper limit on how many subset-free triggers a language might have.)

target language *regardless* of how close to, or far from, the target it was before. Although there can be no direct parallels between a formal model of learning procedures and children's performance, this is clearly not reminiscent of normal child behavior; taken literally, it suggests the unlikely picture of a child making little headway and then all of a sudden exhibiting the full richness of the adult (target) language. To make this approach less incongruous with linguistic and psychological views of learning, this prediction of 'all at once' convergence would need to be tempered, e.g., by dividing the grammar into modules, or by focusing on relevant parameters only. But as noted in section 3.1, these interesting possibilities have not yet been fully worked through to a point where we could estimate their likelihood of success.

Thus, there are two sides to the coin. Unless subset-free triggers are available in every target language that learners are exposed to, an SP-compliant LM might permanently and fatally undershoot its target. The alternative, which we regard as empirically less probable, is that all learnable natural languages do have subset-free triggers; then the problem for learning theory would be to reconcile this with what is known of children's normal progress toward the target.

3.2.2 *Undershoot errors due to the periphery*

Here we take note of another class of undershoot errors that could permanently block convergence on the target. These arise in the course of 'forward' learning, and would exist even if SP did not demand retrenchment. In order to better examine this forward-learning problem, we therefore put the retrenchment requirement on hold during this section. Except in this one respect, all of our usual working assumptions remain active.

Under working assumption A2 the only language-specific facts that need to be acquired are housed in the lexicon, where the parameter values are assumed to reside along with whatever words, morphemes, idioms and other idiosyncratic forms must be stored there because they are not projectable from more basic properties of the language. There is no clear consensus among linguists as to whether this collection of idiosyncrasies is subject to significant linguistic constraints or is completely lawless, falling (perhaps by definition) outside the dictates of UG. This is yet another theoretical linguistic issue with profound implications for models of human language acquisition.

Chomsky (1965, and elsewhere) has observed that learning would be facilitated if UG admits relatively few human languages, distributed sparsely in relation to potential inputs, so that each input sentence is compatible with very few hypotheses. Ideally this would be true of subset relations also: they should be relatively few, so that LM would not have to contemplate many subset languages en route to a more encompassing target language.

However, in place of this optimal design, it appears that s–s-related languages may be very densely packed in the human language domain, multiply nested one within another. In the worst case, the domain might be infinitely densely packed and the learner would forever undershoot the target language. This is not a danger if only 'core' languages are considered. In a parametric framework the number of core languages is finite, a simple function of the number of parameters. The density problem escalates, however, when 'peripheral' facts are taken into account, such as constructional idioms with properties that must be stored because they don't follow from core principles and parameters. Clearly this increases the number of languages in the domain, and hence the number of hypotheses a learner may need to consider. Moreover, the density problem can get totally out of hand, as we will show, if peripheral and core sentences can overlap, i.e., if a core sentence in a language with one set of parameter values can be a peripheral sentence in a language with a different set of parameter values.

Studies of the periphery are scarce, and its relation to core syntax is not well understood. Some linguists regard the core/periphery distinction as an artificial regimentation imposed on a continuum of more-general to less-general syntactic phenomena. Others reject the notion entirely, maintaining that there is no need to recognize a special category of eccentric or isolated constructions exhibiting limited patterns or exceptions, since all natural language sentences, if analyzed correctly, can be licensed by fully general syntactic principles in conjunction with standard lexical entries. Progress in this direction (see, for example, den Dikken 2005) is very welcome, but it could affect the scale of the learning problem only if it thereby reduced the total number of possible languages nested one within the other. Otherwise all the issues discussed below would remain, even if they should be described not as competition between core and periphery but as competition between broader and narrower core generalizations. For expository simplicity in what follows, we will presume a sharp core/periphery distinction, but without intending to beg the theoretical linguistic issues; for example, our conclusions do not rest on the assumption that the periphery is totally unconstrained by UG. Under either formulation, the important observation is that LM cannot first acquire a core language with big broad generalizations (as per parameter setting) and then tack onto it a fringe of oddities and exceptions. Rather, exceptions and syntactic idioms are often located *in between* one core language and another (Fodor 1994). This places them directly in the path of an SP-compliant learner, and significantly increases the learner's workload.[28]

[28] This wouldn't be true if sentences falling under core generalizations could be easily distinguished by learners from genuinely peripheral constructions (e.g., by prosodic cues) so that the latter could be kept clear of the parameter setting process. Some possibilities for recognizing peripheral constructions were examined by Fodor (1994) and rejected. Per-

The class of possible languages consists of all combinations of one of the core languages plus a (possibly empty) set of peripheral constructions. Let us call a language with a non-empty set of peripheral constructions a *peripheral extension* of the largest core language it contains (assuming for convenience that the latter is unique). Every core language is a subset of all its peripheral extensions. Now suppose that one core language (L_a) is a subset of another (L_z), their grammars differing with respect to the value of just one s–s parameter. Some peripheral extensions of L_a may be subsets of L_z. This would be the case if the peripheral extension languages contain subgeneralizations or isolated instances of a phenomenon which is realized as a broad general pattern in L_z. For example, Fodor (1994) cites various eccentric instances of locally-bound pronouns,[29] where L_z is a core language with a parameter-setting that permits locally-bound pronouns in all contexts. SP forces the learner to traverse *all* such languages en route between the two core languages: starting from L_a, it would have to disconfirm every peripheral extension of L_a that is a subset of L_z before being allowed to contemplate L_z.

From the point of view of the learner, this would present itself as follows. Suppose LM is currently in L_a and the target is L_z. LM hears an input sentence (i.e., a sentence of L_z). As long as this sentence is of a type that could be peripheral, SP *requires* LM to tack it onto L_a as a peripheral extension of L_a, rather than treating it as a trigger for hypothesizing L_z. This could happen repeatedly: every sentence of L_z as it is encountered would be tacked on to L_a as a peripheral construction.[30] Hence no sentence of L_z would serve as a trigger for L_z's distinctive parameter value.

There are two subcases to consider. If there is a finite number, n, of sentences in L_z but not in L_a, then the language (the sentence set) L_z would eventually be acquired, though the generalization represented by L_z's parameter setting would not, since the sentences that would have fallen under it would instead be stored individually by LM as n separate idioms.[31] Hence,

---

ipheral constructions do not in general have any superficial mark of their special status, nor are they withheld as input until after the child has acquired the core. Syntactic idioms and oddities abound in language addressed to (and used by) young children. Thus, tracking frequency of occurrence in the input would not help learners distinguish exceptions from triggers for broad generalizations.

[29] An example cited by Fodor (1994) is the English sentence *I'm going to whittle me a walking stick*. Clearly this is not a one-of-a-kind idiom like *kick the bucket*, but is part of a minor generalization in English (cf. *I'm going to make me a sandwich*; *He should take him a long hot bath*), though a core generalization in other languages.

[30] Of course, the SP-driven retrenchment imperative (which we have set aside temporarily) makes it very unlikely that LM would ever get to the point of hypothesizing L_a plus more than one peripheral construction, since it would have to give up all other peripheral constructions it had previously added to L_a each time it added a new one.

[31] Note that this assumes that LM creates individual sentence idioms rather than consolidating sentences into a semi-general constructional idiom. The latter sounds plausible enough as a psychological hypothesis, but SP construed strictly as (d′) would not allow LM

that parameter value would play no part at all; learners could not employ it, so it might as well not exist. (In principle, a learner could adopt L_z once it had encountered all n sentences, but a memoryless LM could not know when it had encountered them all.) Clearly, this effect of SP is diametrically opposed to the presuppositions of most linguistic theories, whether they are cast in terms of parameters or not. SP favors ad hoc and uneconomical descriptions over the broad generalizations that linguists value: if the periphery always took precedence over the core for learners, natural language grammars would be nothing but immense repositories of unrelated facts. The second possibility to consider is that L_z's parameter value adds an infinite number of sentences to those in L_a (e.g., it permits an option in an infinite number of L_a's sentences, such as the option of either Wh-movement or Wh in situ where L_a has only the latter). Now if LM adds each such sentence of L_z, as it encounters it, to the periphery of its grammar for L_a, then even the sentence set equivalent to L_z will never be attained; it will continue to recede into the distance, however much input LM receives. Thus, convergence on L_z is impossible.

This problem might be tamed by imposing a finite cap, m, on how many sentences the periphery of a natural language could contain.[32] Then the first m sentences that LM encounters which lie between L_a and L_z would still be analyzed as peripheral extensions of L_a, but on hearing an $(m+1)$th instance, LM could leapfrog from its current language (a peripheral extension of L_a) directly to L_z, regardless of how many further peripheral extensions of L_a might otherwise have intervened in the absence of any size limit on the periphery. Thus, L_z would be acquirable. But it would be acquirable at the expense of making some 'en route' peripheral extensions unacquirable. If m were imposed by UG, the unacquirable languages would not qualify as legitimate natural languages, so this leapfrogging would not constitute a learnability failure. However, a limit such as m is not the kind of constraint that would typically be thought of as dictated by UG, unless possibly it were to result indirectly from a length limit on peripheral sentences (given that idioms longer than one clause, or two at most, probably don't exist). Alternatively, the bound m might not be UG-imposed but a *performance* limit on the size of the grammar a learner is capable of storing. In that case,

---

to posit any constructional idiom that would result in a wider set of sentences than if each input sentence had been stored individually.

[32] In principle there could be an infinite number of sentences in the periphery of a language. Though they couldn't be listed individually, they could result from a constructional idiom containing variables that permit recursion. However, as noted in fn. 31 above, SP in the form of (d′) would prevent LM from positing any such construction; it would require every input sentence to be stored individually. Thus it follows from strict adherence to SP that in practice there could be only a finite number of sentences in the periphery of any learnable language. However, without a *specific* pre-established cap such as m, L_z would still be unattainable.

some genuine UG-compatible non-core languages might be unlearnable, e.g., any UG-compatible language consisting of L_a plus $(m+1)$ peripheral constructions. Suppose such a language were the target. If LM had memory capacity for only m peripheral facts, then even if it were designed to obey SP in general, on hearing the $(m+1)$th sentence not in L_a it would have no choice but to skip over the target language and hypothesize the core language L_z instead. This would be an inadvertent SP violation. For all we know, human learners do sometimes inadvertently contravene SP in extreme cases where SP demands grammars whose complexity exceeds psychological limits. This would not result in detectable language pathology, since if nobody could acquire the beyond-capacity language with $m+1$ peripheral constructions, then nobody would be exposed to it as a target language.

A linguistically more interesting solution to the periphery problem would stem from a limit not on the number but on the *nature* of peripheral constructions. That is, UG might require some constructions to be treated as core by learners. Something about their structure would rule them out as idioms. Thus, not every sentence in between L_a and L_z would be ambiguous between core and periphery, and those which are unambiguously core could set the core parameters without violating SP. A possible instantiation of this is *UG-designated triggers* for parameters. As outlined by Fodor (1994), one or more triggers would be designated by UG for each (non-default) parameter value and would always trigger that parameter value when encountered by LM. From this it would follow that a UG-designated trigger could never appear as a syntactic idiom in the periphery of any language. Other sentences (not designated triggers) licensed by the very same parameter value would not trigger that value, and so they *could* occur as syntactic idioms in other languages. There are linguistic and learnability benefits here. For learnability: UG-designated triggers for core parameters would permit even a strictly SP-compliant LM to move directly from a current language (e.g., L_a, or one of its peripheral extensions) to a broader core language (e.g., L_z) when the designated trigger for that core language was encountered, without spending time on 'en route' languages but also without sacrificing the ability to acquire an 'en route' language if that were the target (in which case the designated trigger would not occur in LM's input). Thus, designated triggers can mitigate the extreme conservative influence of SP, and permit some rapid learning of major generalizations.[33] For linguistic description, the designated triggers theory has the merit of tolerating a substantial overlap between core

[33] Designated triggers are not identical to subset-free triggers but they are related. Designated triggers resolve core–periphery ambiguities, but not core–core ambiguities. If L_z has no designated trigger, it can have no subset-free trigger, because every sentence of L_z is also in another and smaller language: either in L_a or in a peripheral extension of L_a. If L_z does have a designated trigger, it may or may not have a subset-free trigger. The designated trigger is a sentence that is in L_z but not in any of the peripheral extensions of L_a, so SP cannot give the peripheral extensions priority over L_z. Hence the peripheral extensions

and peripheral items across natural languages. It does exclude some core sentence types from the periphery, but only as many as there are parameters in UG. From the perspective of SP, it maintains the intuitive truth that languages with broad core generalizations are 'good' languages for learners, despite SP's insistence on extreme conservative learning.

To summarize: Even if the retrenchment problem could be solved, undershoot errors could occur as the result of LM's imposing a peripheral analysis on an input sentence that has a core derivation in the target grammar. If the density of peripheral-extension languages is high in the natural language domain, this would cause serious slowdowns in attaining the core languages and failure to capture true generalizations. If there is no bound on the number of sentences that can be ambiguous between core and periphery, learning could fail entirely. Matters only become worse if we now add back into LM the SP-driven need for retrenchment at every step. In that case, as LM moves forward from L_a by adding a peripheral construction to it, it should give up all other peripheral constructions it had previously added to L_a, and would then have to re-acquire those sentences later.

At a more general level of assessment, we note that a model of core acquisition need not itself cover acquisition of the periphery (it might leave that to some other learning mechanism, even if that is less parsimonious than a single learning mechanism for both), but the procedure for core acquisition must not entail the impossibility of acquiring the periphery or vice versa. Though this sounds like a needless worry, we have seen here that under certain conditions it could occur. On one hand, acquisition of some genuinely peripheral constructions would be impossible if parameter setting were modeled as an 'automatic' process that occurs blindly (i.e., disregarding SP), triggering a core generalization on encounter with any sentence that could be core-derived in some language. On the other hand, rigorous adherence to SP can make acquisition of broad (parametric) generalizations impossible. Designated triggers, if they can be substantiated in natural languages, would offer an escape from this choice between two evils. Since designated triggers are by definition sentences that resist peripheral analysis, they would allow parameter setting to proceed as efficiently as if there were no periphery.

3.3 *Pinpointing the source of problems*

To this point, we have been playing out the combined consequences of working assumptions A1–A12, and attempting to control the damage as problems emerged. Though SP has been thought of as the guardian of learnability, we have found that respect for SP is not well-rewarded. To do its job it must be very strictly defined, entailing (d′) for incremental learners. But

---

drop out of consideration, and L_z will have a subset-free trigger just in case it would have had one if there had been only core languages in the domain.

this can do harm as well as good: undershoot errors due to SP-compliance may be as damaging to learnability as the overshoot errors that would result from non-compliance. In this section we have seen that: (i) when LM makes some forward progress, SP insists on backward steps with respect to other properties of the language hypothesized, because they might not have been warranted by previous input (the retrenchment problem); and (ii) in any case, SP limits forward progress to very tiny steps, because input sentences might be constructional idioms or exceptions (the periphery problem). Either of these consequences of SP could slow learning down or sabotage it altogether, unless some feasible defensive mechanisms can be identified.

Though we have surely not exhausted all possibilities within these working assumptions, we move now to the alternative research strategy of stripping them off one by one, to find out which of them was responsible for the problems we have observed. To do this systematically is too vast a task to undertake here, so we will illustrate by withdrawing just one of the twelve. To our eye, the most likely culprit is the memorylessness working assumption A5. Our next move, therefore, will be to see what can be gained by assuming that human learners do have memory either for past inputs and/or for past hypotheses.

4. Reducing the cost: adding memory to the learning model

At center stage in what follows is the relation between incrementality (working assumption A6) and memorylessness (working assumption A5, repeated here for convenience). Up to this point we have treated these as if they were inseparable, but that is not so. Of the two, it is memorylessness that has the greatest impact on learnability.

A5. The learning mechanism is memoryless in the sense that at any step it has no recall of its prior input or its prior grammar hypotheses; it knows only the current input sentence and its grammar hypothesis immediately prior to receiving that input.

A6. Learning is incremental, in the sense that the learner hypothesizes a grammar (perhaps partial, perhaps unchanged from the preceding one) after each encounter with an input sentence; as assumption A5 implies, target language sentences cannot be accumulated in memory and subsequently compared in order to extract generalizations.

By our understanding of these terms, memorylessness entails incrementality: a learner without memory for past learning events cannot afford to wait before formulating its hypotheses, so it must make a decision after each encounter with an input sentence. But incrementality certainly does not entail memorylessness: LM could formulate grammar hypotheses in tempo

with the arrival of input sentences whether or not it could recall past learning events.

Incremental learning is generally depicted as contrasting with a 'little linguist' (hypothesis-formation-and-testing) approach to language learning, in which data are gathered together over time and examined for regularities. Incremental learning tends to be associated with parameter setting, particularly if that is viewed as triggering, while the data-gathering approach is more often associated with rule construction. But these pairings are not invariable: Wexler & Culicover (1980) proposed an incremental rule learner, while Valian (1990) proposed a non-incremental parameter setter. In fact, incrementality *per se* has very little bite. Though important as a consequence of memorylessness, as an independent premise it has no discernible effect when other assumptions (e.g., greediness or the error-driven requirement) are held constant: a non-incremental learner is not so different from an incremental learner except that it may change its mind less frequently. (Possibly, the stricter timing that incrementality imposes on hypothesis formulation could affect learning outcomes if hypothesis *selection* principles are applied on all and only those occasions on which LM actively adopts a new hypothesis.)

Rather, incrementality has been a focus of interest in psycholinguistics not so much for itself but precisely because it permits tight bounds on the memory resources that must be attributed to learners (and on the complexity of the computations they must engage in; see below). Each input sentence is absorbed immediately into the current grammar, which has to be in memory in any case for purposes of language comprehension and production. It is the grammatical implications of an input sentence that an incremental LM stores, not the sentence itself. Sadly, what section 3.1 documented is that the mental grammar is not a reliable device for storing the grammatical implications of past input, because SP requires the hypothesized grammar to be completely refreshed each time it needs updating at all. This is retrenchment, and it is highly destructive to learning. Retrenchment might be limited if a learner could remember past input sentences; then the smallest language compatible with its known input would not, after all, be so small. Or LM might remember which grammars it had previously rejected so that even the retrenchment requirement couldn't force it to keep re-hypothesizing them. Thus, learning theory has a choice as to *where* memory should be added to LM in order to escape the problems of 'purely' incremental (memoryless) learning. The alternatives are as we anticipated in SP(d), the memory-sensitive formulation of the subset principle (repeated here).

SP(d): When LM's current language is incompatible with a new input sentence *i*, LM should hypothesize a UG-compatible language which is a smallest superset of *i and all prior input sentences retained*

> *in its memory*, excluding any language *recorded in memory as having been disconfirmed by prior input*.

SP(d) delineates the extent of retrenchment in terms of the input sentences that LM knows it has already encountered, and the grammars it knows it has already disconfirmed. Enriching LM's memory for either of these sets could block the unwelcome entailment from SP(d) to the excessively restrictive (d′).

Our next topic of investigation, therefore, is whether undershoot errors can be reduced by increasing LM's ability to recall past inputs, and/or to recall past grammar hypotheses. Both types of memory have been proposed in the literature. Within the psychological tradition, hypothesis-testing models traditionally assume the former. In a computational linguistic context Brent (1996) advocated the latter. Gold (1967) assumed both. We will not review past proposals here. Instead, we will consider each form of memory in turn to assess specifically whether it could free SP to do its essential work.

4.1 *Memory for prior input*

Imagine that LM could recall the last four input sentences prior to the current one. Then SP(d) directs LM to hypothesize a smallest language compatible with the five sentences it has knowledge of. This is likely to be larger than a smallest language containing only the current sentence. Some retrenchment might still occur, with loss of knowledge that was acquired from input prior to those five sentences, but the extent of the loss would be reduced compared with memoryless learning. Also, there would be a higher probability that the target language has a subset-free trigger when a subset-free trigger can consist of five sentences rather than one. More languages would be learnable.[34] These benefits would increase with the extent of memory: a collection of 50 or 500 or 5,000 sentences is more likely to uniquely trigger one target language than is a collection of 5. At some point, though, the curve of increasing benefits would likely cross a curve of increasing memory overload. The burden on memory is bound to grow as learning proceeds, if LM attempts to store thousands of input sentences.

The feasibility issues here are similar to those faced by a non-incremental, hypothesis-testing learning model, which have been accorded surprisingly little attention in the literature. One issue begging to be investigated is the

[34] The formal learnability literature contains theorems which appear to contradict the claim that more languages could be learned with greater memory for past input (e.g., see Osherson et al. 1986, section 4.4.1), but there is no conflict with our observations here. We are addressing not the learnability of language domains (i.e., what classes of languages can ultimately be learned regardless of practical considerations), but the psychological feasibility of learning natural languages. Nevertheless, insights gained from formal learning-theoretic investigations can prove useful in psycholinguistic modeling. For example, in section 4.2 we borrow the concept of an enumeration, which figures in many formal proofs and which is a form of memory for disconfirmed grammars.

computational cost of fitting a grammar to an ad hoc collection of many sentences or to a body of facts extracted from those sentences. Our primary concern here, however, is memory load. How much of the language could a child plausibly store in *addition* to the language facts encoded in the grammar that he or she has hypothesized so far? Children may have memory capacity sufficient for the last few sentences of a conversation. They might even save up all the sentences they hear during a day and spend late-night hours in the crib reviewing them (Weir 1962). But a collection of input sentences doesn't constitute a natural class, so it doesn't lend itself to summarization to lighten the load. So even if memory span increases as a child grows older, it is unlikely to be able to keep up with the total accumulated language experience.[35] This suggests that while no memory at all for input may inhibit learning, memory for the total input sample is not the cure.

But perhaps LM doesn't need to preserve it all. What is the minimum that would need to be retained in memory in order for LM to apply SP but avoid wasteful SP-driven retrenchment? We have seen that the knowledge LM has acquired must be renewed at every grammar change, so memory must preserve it or it will be lost. But it doesn't follow that every one of the past *n* inputs must be accessible in memory. LM might extract useful information from input sentences, such as tree configurations or abstract 'cues', and then discard the sentences. However, since input sentences can be ambiguous, any such facts drawn from them could be just as much in need of subsequent revision as parameter settings are. So it seems that what is remembered must stay very close to LM's actual raw data. Other ideas would be very welcome, but we will suppose that it is essentially input sentences themselves that are stored. Fortunately, an exact recording of the total input stream is not called for. Some savings are possible. Repeated sentences needn't be stored. Repeated sentence *types* needn't, either, once a learner can recognize those types. All that LM will want to access later is the relatively few sentences (sentence types) that could make a difference between one grammar hypothesis and another. In P&P theory, these are the triggers for parameters (setting aside the periphery now). If triggers are innately specified, they could be recognized as they occur and retained in memory. If not, the valuable input sentences could be identified by LM as those which caused a change in

[35] Learning models which register frequencies of occurrence presuppose considerable memory. If the frequency of some item (a word, a collocation, a sentence) is mentally recorded, then that item must itself be mentally stored (albeit perhaps in some coded form) in association with its frequency score. While the assumption of storage for lexical items (including syntactic or semantic idioms, as in the periphery) is unproblematic, storage of productive word combinations falls, in our judgment, beyond plausible resource limits. For this reason, the future of statistical learning as a component of a psychological model of syntax acquisition will depend on whether the frequencies of phrases and sentences prove to be reducible to the frequencies of a more limited number of construction types, or to the abstract principles which define them, or possibly to some dynamic encoding of them. We have begun to consider this issue in Kam et al. (2005).

its grammar hypothesis when they occurred.[36] These are the sentences worth storing in memory for future reference. Basing a grammar hypothesis on the last five or fifty sentences may be better than basing it on just one, but basing it on these informative trigger sentences would be better still. When retrenchment occurs these sentences would not be erased from LM's hypothesized language, so all the sentences they normally trigger would remain as well. Thus, if LM is to be empowered with memory for input, selective memory for triggers would seem to be optimal: it would provide the greatest bulwark against SP-enforced retrenchment at the smallest cost in memory. In future work the gain from adding memory for past triggers needs to be evaluated and quantified. Related issues are given formal treatment by Lange & Grieser (2003 and references there; also Jain et al. 1999).

To summarize: Giving LM the capacity to remember some past inputs appears to be both beneficial and psychologically feasible as long as LM is judicious about which sentences it stores. This notion of *selective memory for input* has not previously been discussed in the psycholinguistics literature. Though it is unconventional and in need of more study, as a way of loosening up strict incrementality it seems to have some potential.

4.2 *Memory for prior hypotheses*

SP(d) relates LM's selection of a next hypothesis not only to its memory for input but also to its memory for language hypotheses that have been disconfirmed. We turn our attention now to the latter. Putting aside any memory capacity for input, we now consider whether retrenchment could be safely minimized if LM could recall which grammars it has already tried out and rejected. First we set out the benefits of such a system and evaluate its compatibility with psychological constraints. Then in section 4.3 we will consider obstacles to the efficient exploitation of this type of memory.

With memory only for previously rejected hypotheses, the language that LM should guess on hearing sentence *i* would be a smallest *non-falsified* language containing *i*. Over time, an increasing number of smallest languages in the domain would be falsified. They would not need to be visited again, so LM could hypothesize progressively richer languages without clashing with SP. Note that this is more valuable than merely saving LM from wasting time on 'accidental' retesting of a previously tried hypothesis. With retrenchment in full force, we know that LM would return unproductively

[36] This criterion for determining which items to retain in memory would be a rough and ready one, since some sentences that did not trigger a hypothesis change when they were encountered may carry significant information in the context of a different hypothesis later on, and vice versa. Also, if a trigger was ambiguous when it was encountered, it would still be ambiguous if stored and made use of later. Nevertheless, even an approximate culling of 'most valuable input sentences' could be useful in limiting memory cost while providing a richer database for current hypotheses.

to the *same* few smallest languages over and over again. Memory for past hypotheses allows LM to break out of this trap. However, despite its advantages, the costs of this solution could be extreme. A complete tally of failed hypotheses, one by one, would place an unreasonable strain on memory: LM would need to commit as many memory cells as there are possible languages, e.g., a billion memory units for a language domain defined by 30 binary parameters. Understandably, a list of all disconfirmed hypotheses is *not* standardly assumed in psycholinguistic models of acquisition.[37]

On the other hand, a familiar component of computational models since Gold (1967) is an *enumeration* of all possible languages, i.e., a complete ordering of all the languages, which determines the sequence in which LM hypothesizes them. As discussed above (section 2), an enumeration could play a significant role in a psychological learning model by providing LM with the s–s information necessary for applying SP. Subset languages precede all their supersets in the enumeration, so both undershoot and overshoot errors are avoided (in a finite domain) as long as the learner proceeds through the languages in sequence, never returning to a previous (disconfirmed) one, and not moving on to a later one until it has determined that intervening ones are incompatible with the data. The class of learnable human languages would consist of all and only the languages which precede their supersets in the enumeration. A partial ordering is sufficient to encode s–s relations, but a *total* ordering, as is commonly assumed in the Gold paradigm, has three additional benefits for a learner lacking other types of memory: (a) it provides an economical means of blocking excessive SP-driven retrenchment; (b) it prevents unproductive loops in which (regardless of retrenchment) LM would hypothesize the same languages repeatedly (see fn. 22); (c) it eliminates the need for subset-free triggers. We consider these in turn.

(a) If LM has access to the enumeration, then it knows that it has already disconfirmed all languages which precede its current hypothesis in the enumeration and has tested none which follow it. The current hypothesis thus serves as a pointer. The enumeration is innately given, so its demand on on-line resources is negligible. The pointer is in working memory, since it shifts as learning proceeds, but it is a very simple mnemonic, and it is cost-free since LM must in any case be assumed to know what its current grammar is. Thus LM is able to avoid retrenchment to any previously falsified hypothesis, without having to record falsified grammars one by one.

[37] Opinion on memory for disconfirmed grammars is divided. Brent (1996) advocates it. Pinker (1979) vehemently opposed one implementation of it (an enumeration; see below). Yang (2002, chapter 2) considered it before moving to a less costly, but also less precise, recording system in terms of the success or failure of parameter values rather than whole grammars; see also Fodor (1998a) for a parameter-based memory system.

(b) SP permits LM to move freely from one intersecting language to another; since no overgeneration errors can result, SP does not require them to be tested in any particular sequence. However, LM does need to know when it has checked them all, so that it won't continue to test and re-test them repeatedly. A shift from a partial grammar ordering to a total grammar ordering as in an enumeration can assist with the necessary record-keeping. In the total ordering, languages that don't need to be ordered with respect to each other for SP purposes will have an arbitrary but fixed order imposed on them. Now LM will explore this collection of intersecting languages in a specific sequence, and when it has worked through them all, it can move forward to new (superset) hypotheses.

(c) An enumeration also makes it possible to acquire a language which lacks any subset-free triggers. As observed in section 3.2, for an incremental learner *without* an enumeration, such languages are not reliably learnable: learning failure can result from overgeneration if LM does not obey SP, and from undergeneration if it does. However, if the learner has an enumeration of a finite domain of languages, formal results show that convergence is not impaired by lack of memory for input (though it can be slower than for a comparable learner with memory for input; see Bertolo 1995b). It is clear why this is so. A subset-free trigger is not needed for learnability if an enumeration is available to keep track of grammars already tested. LM can safely adopt a language however many subsets it has, once it has tested all those subsets and rejected them. With an enumeration, there need not be any *one* input sentence (any subset-free trigger) that rules out all the subset hypotheses.

Thus, an enumeration brings learning benefits of several kinds, all of which can channel an incremental learner into making steady progress towards the target grammar. Is it plausible, then, to suppose that an enumeration is part of the standard equipment by means of which children acquire their native language? This is not out of the question. If the number of languages is finite, the enumeration might be inscribed as such in the brains of human infants. Or, if there is some predictable pattern to the ordering, infants might be innately equipped with a productive mechanism for projecting the enumeration. These two possibilities (an innate list of languages, or a general metric for ordering them) are mechanisms we have already had occasion to consider as providing the *partial* ordering needed for SP-compliance. They correspond to alternatives (ii) and (iii) of section 2. Expanding that partial ordering into a full ordering is conceptually a minor amendment. It also appears to carry little practical cost, at least if the enumeration is projectable via general principles (i.e., (iii) rather than (ii)). In section 2.2 we considered two prioritization schemes based on a parametric format for grammars: the Simple Defaults Model and the Ordered Defaults Model. The latter (with a default value for each parameter together with a priority

ordering of the parameters) could easily be made to yield a total ordering of languages, given some simple algorithm to systematically read grammars off the parameter vector. Best of all, of course, would be a systematic ordering that is also linguistically and psycholinguistically motivated. This should be high on the research agenda for acquisition theory *if* the basic idea of enumeration can be shown to pass muster.

In sum: Knowledge of disconfirmed hypotheses would supplement an incremental learner's information base so that it could learn cumulatively; it would not suffer from repeated retrenchment to extremely small languages. Moreover, a record of falsified grammars may not be unrealistic for human learners, since an innate enumeration of grammars serves as a record-keeping system which converts knowledge of the current grammar hypothesis into knowledge of all disconfirmed grammar hypotheses. And though parameters were originally conceived as unordered, there have been some arguments unrelated to SP in favor of ordering them, as noted in section 2.2.

Some psychological analogue of the technical concept of an enumeration would thus appear to be an effective way to make the consequences of SP tolerable for incremental learning.[38] Though not a familiar concept in psycholinguistics, it is not really a foreign one. In fact, as noted in section 2.1, an enumeration is essentially equivalent to an evaluation measure. Our discussion has thus come full circle, to the question of whether a psycholinguistically satisfactory evaluation measure for syntax acquisition can be defined. If it can, the problems that arose for SP in memoryless learning appear to be solved. Further, if that evaluation measure can be shown to be *necessary* for natural language acquisition, it would provide support for linguistic nativism, as Chomsky (1965) anticipated, since the enumeration could not itself be learned. Involvement of an enumeration (whether listed or projected) implies an innate preparation for language acquisition, regardless of whether any particular language facts (e.g., there are nouns and verbs) or any particular linguistic principles (e.g., Subjacency) are innately known. In this respect, the Gold paradigm and the Chomsky paradigm concur.

It will be interesting to compare the efficiency of an enumeration and the efficiency of input memory, as antidotes to retrenchment. We cannot say at present which offers a better fit to child acquisition data. On technical grounds the enumeration approach has the advantage, since it addresses both of the major issues for SP-implementation. It provides a means for representing s–s relations (section 2), and also mitigates the adverse consequences of applying SP on-line (section 3), while memory for input sentences

[38] Note, however, that like memory for past inputs, an enumeration leaves the periphery problem untouched. All peripheral languages would need to be included in the enumeration, and something like designated triggers would still be needed in order to skip past them to a core language. Though we won't stress this in what follows, the presence of many peripheral languages in the enumeration would magnify the practical challenges (see below) in using it efficiently.

does only the latter. Also, an enumeration might be computationally easier to employ on-line, since LM need only find a grammar that licenses the current input sentence (as in 'pure' incremental learning), rather than devising a way to license a whole collection of sentences. It must be emphasized, though, that a *total* grammar ordering is essential to reap the bookkeeping benefits of the enumeration approach, so if that is not consistent with the facts of child language acquisition, this solution loses interest. Moreover, whatever merits it may have, the enumeration/evaluation metric approach comes with a high price tag, as we now discuss.

4.3 *Enumeration and input decoding*

Though a mainstay of computational learning theory, the concept of an enumeration of languages/grammars has not figured prominently as a psychological hypothesis. This may be because the devastating effect of SP on incremental learning has not previously been noted,[39] so the ability of an enumeration to keep that in check has not been valued. But enumeration may also be unappreciated because it appears to demand an unrealistic amount of on-line processing of input sentences.[40]

Some familiar learning algorithms (such as those of Gibson & Wexler 1994, Briscoe 1999, 2000, Yang 2002) respect human resource limitations by allowing LM to try out only one new grammar per input sentence after its current grammar fails. For example, Gibson & Wexler's TLA attempts to parse the sentence with the current grammar; if that fails it selects another grammar (by randomly changing one parameter value) and tries to parse the sentence with that; if that fails, it stops trying and retains the current grammar. But let us now imagine a learner just like this except that its choice of a new grammar to test against the current input sentence is determined by an innate enumeration of grammars: it chooses the grammar that immediately

[39] Wexler & Culicover (1980: 525, chapter 3, fn. 16) come closest. They observe the disastrous effect of a simplicity metric on incremental learning: 'The selected grammar will simply continuously shift, dictated by only one datum.' The argument they give makes reference to an enumeration, which they recognize as equivalent to an evaluation measure. Throughout the chapter they present several objections to learning strategies that rely on an enumeration. Though they also consider some antidotes, their ultimate vote is apparently against enumeration. But it should be noted that their own *degree-2* acquisition model did not have to confront SP problems because other assumptions provided it with indirect negative evidence.

[40] Pinker's (1979) objection against importing a Gold-type enumeration into a psychological model was on grounds of efficiency: the 'depressing result is the astronomical amount of time that the learning of most languages would take. The enumeration procedure … exacts its price: the learner must test astronomically large numbers of grammars before he is likely to hit upon the correct one' (p. 227). Pinker summarizes (p. 234): 'In general, the problem of learning by enumeration within a reasonable time bound is likely to be intractable'. As we discuss below, if an enumeration is to be usable, there must be efficient ways of moving through it.

follows its just-failed current grammar in the enumeration; if that fails, it reverts to the current grammar. We will call this learner the *Enumeration-LM*. The TLA has a free choice in selecting its next grammar hypothesis, within the limits of the SVC and Greediness; it is not constrained by SP (and *a fortiori* is not constrained by an enumeration). By contrast, the Enumeration-LM has no freedom of choice with respect to its next hypothesis, but it does respect SP.

The Enumeration-LM performs very poorly. This is because in many cases there is more than one grammar in the enumeration between the current grammar and the next grammar that is compatible with the input (henceforth: *the next compatible grammar*). In such a case, when the Enumeration-LM tries out the grammar that immediately follows its current grammar in the enumeration, that grammar will fail and the learning trial will be over; the current grammar will be retained unchanged. The input sentence has thus contributed nothing. Even if it were a highly informative sentence, capable *in principle* of moving the learner many grammars ahead in the enumeration, it is powerless to do so. This wastage of potentially informative input is a familiar drawback of trial-and-error learners (see Sakas & Fodor 2001). But for the Enumeration-LM with a limit of one new grammar-test per input sentence, it presents two quite specific problems.

One problem is that the learner's current grammar cannot now serve as a 'pointer' to mark the division between grammars already disconfirmed and grammars not yet tested. When the Enumeration-LM tests the next grammar in the enumeration and it fails, its grammar does not change, so the pointer does not move along the enumeration. Hence no record is kept of the fact that this grammar was checked and disqualified. So the same wrong grammar will be tried on the next input sentence and the next, and no progress will be made at all until a sentence occurs which this wrong grammar does successfully parse; then the pointer can be moved one step forward. Thus, the Enumeration-LM will not converge on the target unless by improbable good fortune the target language (which supplies the input) happens to contain at least one sentence in each of the languages that lie between the current grammar and the target grammar. If it doesn't, the Enumeration-LM will at some point become permanently stalled, unable to proceed further (i.e., it would be in a local maximum). But there is a straightforward remedy for this. If the greediness assumption (working assumption A8) is given up, the Enumeration-LM could adopt a grammar it is testing regardless of whether that grammar passes or fails the test. By the error-driven learning constraint (working assumption A7; cf. section 1.2 above) that grammar would be being tested only if the previous grammar had failed on the current input, so there would be no risk that this regimen would bypass a valid grammar. And it would ensure that LM's current grammar is always at the dividing line between disconfirmed grammars and untested grammars, as desired. An alternative approach would be to retain greediness but assume

that the Enumeration-LM has a purpose-built mental pointer to record the furthest grammar in the enumeration that has been tested, regardless of whether or not this coincides with the current grammar. Either way, this is a problem that can be overcome.

The second consequence of limiting the number of grammar tests per input sentence is more difficult to defend against. With or without greediness, and with or without a separate pointer, there is no escaping the necessity for the Enumeration-LM to test *every* grammar that appears prior to the target grammar in the enumeration. On present assumptions that will consume at least one input sentence for each grammar tested. Learning will be fast in some cases but extremely slow in others, the time to convergence being a function of how far along in the enumeration the target grammar is located. If indeed all of a billion or so UG-defined languages are learnable by humans, then in the worst case all but the target would have to be disconfirmed, and a corresponding number of input sentences would be needed to do so (see fn. 40). This would be a steep cost for LM to have to pay in return for avoiding constant retrenchment. Is there a more efficient way of utilizing an enumeration?

Note that the Enumeration-LM, as we have characterized it, does nothing *except* keep a record of which grammars it has disconfirmed, one by one in the sequence (whether the sequence is projected or listed). It has no strategies for zooming in on where the correct grammar is located in the enumeration, so it cannot afford to skip any intervening grammars. In the Gold paradigm, LM gains the benefits of an enumeration without suffering this disadvantage. Gold's learner moves forward on each trial to the next *compatible* grammar, skipping over intervening grammars however many of them there might be. How this could be achieved was not discussed since psychological resource limits were not at issue; the goal was to set bounds on learnability in the limit, rather than to evaluate feasible learning by resource-limited human learners. But for our purposes this is precisely what we need to know: Can a plausibly resource-limited model be devised that will skip to the next compatible grammar? One might contemplate stretching the one-grammar-test-per-sentence limit to two or three per sentence, but there is no reasonable extension that would allow an Enumeration-LM always to reach the next compatible grammar regardless of how far ahead it is. An alternative would be for LM to have advance knowledge of which grammars are compatible with any given sentence, so that no on-line tests would be necessary at all. Then LM could just consider those candidates, and select from them the one that appears next in the enumeration; it would move in one step straight to the next compatible grammar. At the beginning of section 3 we conjured into existence an imaginary database that would supply LM with useful knowledge of all sorts, including knowledge of which grammars could license a given sentence. However, that was not intended as a component of a realistic learning model but was a device for staving off questions about what

LM does or does not know about the language domain. Now we seek real answers to those questions.

Can a human learner, given an input sentence, identify the set of all grammars compatible with it? This is what we have termed 'decoding' an input sentence, as described in section 3.1. It is of interest independently of SP. Recent work has shown that decoding is a very difficult task, but also that it is a major contributor to learning efficiency since it focuses the learner's attention on just the grammars that have a chance of being correct, freeing it from the need to search through the huge domain of all possible grammars. Most models do not have a decoding component, but the Structural Triggers Learner (STL; cf. end of section 3.1) is capable of a limited amount of decoding: for each input sentence it finds *one* combination of parameter settings that could license it, with no guarantee that these settings are the right ones if the sentence is ambiguous. In simulation experiments we have shown that (some variants of) the STL outperform(s) non-decoding learning systems in terms of accuracy and learning speed (Fodor & Sakas 2004). Decoding is also the key to making effective use of an enumeration. A learner that could decode sentences on-line without undue effort could move through an enumeration in just the way that Gold envisaged: in large steps or small ones depending entirely on what the input dictates. When a decoding LM has established the set of grammars compatible with an input sentence, it can go directly to whichever one of them appears first in the enumeration, skipping perhaps hundreds or hundreds of thousands of intervening grammars as it does so. Such a learner *could* reap the advantages of an enumeration/evaluation measure without being slowed down by it; it would move through it just as rapidly as the input affords.

Note, however, that this efficiency is premised on LM being capable of *exhaustive* decoding of an input sentence, i.e., being able to find *every* grammar that could license it. Otherwise, if LM unwittingly overlooked one or more such grammars, it might select a grammar too far advanced in the enumeration, violating SP.[41] Exhaustive decoding would be unnecessary in a domain in which all triggers were guaranteed to be unambiguous, so that

[41] This problem is not exclusive to a parametric framework. A comparable situation arises for rule-based or construction-based grammars: exhaustive identification of *all* candidate grammars is necessary if SP is to select the correct one. Neglect of this requirement goes all the way back to Chomsky's *Aspects*, which proposed to apply an evaluation measure to select the best grammar that could license the current input, though nothing in that theory ensured that the learner would generate a complete set of candidates. Lacking that, an evaluation measure could not reliably impose SP or any other criterion of interest. Clearly, though, what Chomsky envisaged was an initial culling of candidate grammars from which the evaluation measure would select the highest-ranked, in contrast to the inefficiency of first picking the highest-ranked grammar and only then checking it for candidacy, as non-decoding models do. In this respect Chomsky's conception of evaluation was on exactly the right track, and lacked only the crucial (and perhaps unattainable) mechanism for exhaustive decoding.

decoding could stop as soon as LM had identified one candidate grammar; but as we have observed, that is a luxury that human learners cannot count on. We know of no psychologically plausible learning model that does exhaustive decoding of ambiguous triggers. For an STL to do so it would have to be able to execute a full parallel parse of any sentence, but there are reasons to doubt that even adult perceivers can do that, let alone two-year-olds (see Fodor 1998a, and discussion of the Strong-STL in Sakas & Fodor 2001). A psychologically plausible STL performs only serial parsing and finds just one grammar compatible with an input sentence. Depending on details of the model, this grammar may be picked at random or it may be selected by the learner's parsing strategies or other preference principles, but in no case is there is a guarantee that it would be the next compatible grammar in the enumeration, as SP requires.[42] Thus, partial decoding of ambiguous inputs, which is the most that human learners can reasonably be expected to achieve, does not allow LM to make safe progress through an enumeration without having to stop at every grammar along the way.

To summarize: An enumeration is a significant ingredient of many formal learnability proofs, and is close kin to the sort of evaluation measure that has been central to linguistic thinking about acquisition. In section 4.2 we conjectured that significant benefits might accrue from incorporating an enumeration/evaluation measure into a psychological learning model. However, we have now observed that the usefulness of an enumeration depends on the psychological resources supporting LM's movements through it. Non-decoding learners can move through it only very painstakingly, while decoding learners could in principle zoom directly to the optimal hypothesis but only by expenditure of unrealistic parsing resources. Thus despite its promise, there is at present no convincing plan for solving SP problems efficiently by this means.

5. Conclusion

Gold (1967) showed that the Subset Principle (though not by that name) is necessary but not sufficient in the general case for learning from text (i.e., from positive data only). For linguists, SP is more familiar through the work of Baker (1979), Pinker (1979), Dell (1981), Berwick (1985), Manzini & Wexler (1987), Clark (1992), and others. Within the research tradition stemming from Gold's seminal study, these scholars have embedded SP into a rich linguistically-based acquisition theory with aspirations to psychological

[42] A topic for future investigation is whether the parser itself could be made sensitive on-line to the s–s information in the enumeration/evaluation metric, in addition to its usual concerns of Minimal Attachment, etc. If so, even a learner fed by a serial parser would not improperly skip any subset hypotheses in the enumeration.

feasibility. As a result, the penalty for not obeying SP has long been understood. But the costs of obeying SP have not been widely recognized. Clearly they must be addressed if SP is to be built into working models of language acquisition without impairing their ability to meet other standards we want to hold them to: reliability, efficiency, psychological plausibility. In this paper we have drawn attention to problems that emerged as we began to think about the mechanics of applying SP. We have asked: How much does LM need to know about the language domain in order to be in a position to apply SP? How can that knowledge be mentally represented? How complex are the computations by which SP is applied in the selection of hypotheses during learning? What effect does SP have on the speed of learning? What effect does SP have on the chances of eventual convergence on the target? What learning procedures or resources could protect learners against the less desirable effects of SP?

Since the ramifications of these questions go far beyond what can be covered here, we narrowed our sights by adopting some working assumptions, and all conclusions drawn here are necessarily conditional on them. Similar questions can be asked and answered on the basis of other assumptions, within different linguistic frameworks and/or different learning models. It will be of considerable interest to see whether comparable obstacles are encountered, or whether they simply melt away when approached from a different perspective. We have concentrated here on incremental (in the sense of memoryless) learning because it seemed likely to be particularly susceptible to the potentially harmful effects of SP. Indeed we spied an impending crisis, which demanded urgent attention. The realization that SP compliance can cause incremental learning to fail (except in a language domain with unnatural properties; section 3.2.1) was something of a shock, since in the psycholinguistic framework we are working in, both incremental learning and SP tend to be taken for granted as essential aspects of the human language learning mechanism. That they might be ultimately incompatible was not anticipated. We have set out here some ideas for ways in which they might be reconciled. Some seem more promising than others, but we have not found any of them fully compelling. We hope readers will trace out their own routes through the maze of alternatives sketched here, in search of a definitive solution.

REFERENCES

Angluin, D. (1980). Inductive inference of formal languages from positive data. *Information and Control* **45**. 117–135.

Atkinson, M. (1992). *Children's syntax: an introduction to Principles and Parameters theory*. Cambridge, MA: Blackwell Publishers.

Atkinson, M. (2001). Learnability and the acquisition of syntax. In Bertolo (ed.), 15–80.

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

Bertolo, S. (1995a). Maturation and learnability in parametric systems. *Language Acquisition* **4**. 277–318.

Bertolo, S. (1995b). *Learnability properties of parametric models for natural language acquisition*. Ph.D. dissertation, Rutgers, The State University of New Jersey.
Bertolo, S. (ed.) (2001). *Language acquisition and learnability*. Cambridge: Cambridge University Press.
Berwick, R. C. (1985). *The acquisition of syntactic knowledge*. Cambridge, MA: MIT Press.
Berwick, R. C. & Niyogi, P. (1996). Learning from triggers. *Linguistic Inquiry* **27**. 605–622.
Borer, H. & Wexler, K. (1987). The maturation of syntax. In Roeper & Williams (eds.), 123–172.
Brent, M. R. (1996). Advances in the computational study of language acquisition. *Cognition* **61**. 1–38.
Briscoe, E. J. (1999). The acquisition of grammar in an evolving population of language agents. *Electronic Transactions on Artificial Intelligence* **3**. 1–32.
Briscoe, E. J. (2000). Grammatical acquisition: inductive bias and coevolution of language and the language acquisition device. *Language* **76**. 245–296.
Chater, N. & Vitányi, P. (2005, submitted). A simplicity principle for language learning: re-evaluating what can be learned from positive evidence.
Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.
Chomsky, N. (1986). *Knowledge of language: its nature, origin, and use*. New York: Praeger.
Chomsky, N. (1995). *The Minimalist program*. Cambridge, MA: MIT Press.
Clark, R. (1989). On the relationship between the input data and parameter setting. *Proceedings of the 19th Annual Meeting of the North East Linguistic Society* (*NELS* **19**), 48–62.
Clark, R. (1992). The selection of syntactic knowledge. *Language Acquisition* **2**. 83–149.
Crowther, C., Fodor, J. D. & Sakas, W. G. (2004). Does ungrammatical input improve language learning? Presented at the Architectures and Mechanisms for Language Processing Conference (AMLaP-2004), Provence.
Dell, F. C. (1981). On the learnability of optional phonological rules. *Linguistic Inquiry* **12**. 31–37.
den Dikken, M. (2005, to appear). Comparative correlatives comparatively. *Linguistic Inquiry* **36**.
Déprez, V. & Pierce, A. (1993). Negation and functional projections in early grammar. *Linguistic Inquiry* **24**. 25–67.
Dresher, B. E. (1999). Charting the learning path: cues to parameter setting. *Linguistic Inquiry* **30**. 27–67.
Dresher, B. E. & Kaye, J. D. (1990). A computational learning model for metrical phonology. *Cognition* **34**. 137–195.
Felix, S. W. (1992). Language acquisition as a maturational process. In Weissenborn et al. (eds.), 25–51.
Fodor, J. D. (1992). Learnability of phrase structure grammars. In Levine, R. (ed.), *Formal grammar: theory and implementation* (Vancouver Studies in Cognitive Science **2**). Oxford: Oxford University Press. 3–68.
Fodor, J. D. (1994). How to obey the subset principle: binding and locality. In Lust et al. (eds.), 429–451.
Fodor, J. D. (1998a). Parsing to learn. *Journal of Psycholinguistic Research* **27**. 339–374.
Fodor, J. D. (1998b). Unambiguous triggers. *Linguistic Inquiry* **29**. 1–36.
Fodor, J. D. (2001). Setting syntactic parameters. In Baltin, M. & Collins, C. (eds.), *The handbook of contemporary syntactic theory*. Oxford: Blackwell. 730–767.
Fodor, J. D. & Crain, S. (1987). Simplicity and generality of rules in language acquisition. In MacWhinney, B. (ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Lawrence Erlbaum. 35–63.
Fodor, J. D. & Sakas, W. G. (2004). Evaluating models of parameter setting. *Proceedings of the 28th Annual Boston University Conference on Language Development* (*BUCLD* **28**), 1–27.
Frank, R. & Kapur, S. (1996). On the use of triggers in parameter setting. *Linguistic Inquiry* **27**. 623–660.
Gibson, E. & Wexler, K. (1994). Triggers. *Linguistic Inquiry* **25**. 407–454.
Gold, E. M. (1967). Language identification in the limit. *Information and Control* **10**. 447–474.
Hale, M. & Reiss, C. (2003). The subset principle in phonology: why the tabula can't be rasa. *Journal of Linguistics* **39**. 219–244.
Halle, M. (1962). Phonology in generative syntax. *Word* **18**. 54–72.

Hyams, N. M. (1986). *Language acquisition and the theory of parameters*. Dordrecht: Reidel.
Jain, S., Osherson, D., Royer, J. S. & Sharma, A. (1999). *Systems that learn: an introduction to learning theory* (2nd edn.). Cambridge, MA: MIT Press.
Joshi, A. K. (1994). Commentary: some remarks on the Subset Principle. In Lust et al. (eds.), 509–513.
Kam, X. C., Stoyneshka, I., Tornyova, L., Sakas, W. G. & Fodor, J. D. (2005). Non-robustness of syntax acquisition from n-grams: a cross-linguistic perspective. Presented at the 18th Annual CUNY Sentence Processing Conference, Tucson, AZ.
Kanazawa, M. (1994). *Learnable classes of categorial grammars*. Ph.D. dissertation, Stanford University.
Kapur, S. (1994). Some applications of formal learning theory results to natural language acquisition. In Lust et al. (eds.), 491–508.
Keenan, E. & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry* **8**. 63–99.
Keller, F. & Asudeh, A. (2002). Probabilistic learning algorithms and optimality theory. *Linguistic Inquiry* **33**. 225–244.
Lange, S. & Grieser, G. (2003). Variants of iterative learning. *Theoretical Computer Science* **292**. 359–376.
Lust, B., Hermon, G. & Kornfilt, J. (eds.) (1994). *Syntactic theory and first language acquisition: cross-linguistic perspectives* (vol. 2): *Binding, dependencies, and learnability*. Hillsdale, NJ: Lawrence Erlbaum.
MacLaughlin, D. (1995). Language acquisition and the subset principle. *The Linguistic Review* **12**. 143–191.
Manzini, M. R. & Wexler, K. (1987). Parameters, binding theory, and learnability. *Linguistic Inquiry* **18**. 413–444.
Nyberg, E. H., III. (1992). *A non-deterministic, success-driven model of parameter setting in language acquisition*. Ph.D. dissertation, Carnegie Mellon University.
Osherson, D. N., Stob, M. & Weinstein, S. (1986). *Systems that learn: an introduction to learning theory for cognitive and computer scientists*. Cambridge, MA: MIT Press.
Pinker, S. (1979). Formal models of language learning. *Cognition* **7**. 217–283.
Prince, A. & Tesar, B. (2004). Learning phonotactic distributions. In Kager, R., Pater, J. & Zonneveld, W. (eds.), *Constraints in phonological acquisition*. Cambridge: Cambridge University Press. 245–291.
Randall, J. H. (1992). The catapult hypothesis: An approach to unlearning. In Weissenborn et al. (eds.), 93–138.
Roeper, T. & de Villiers, J. (1992). Ordered decisions in the acquisition of *wh*-questions. In Weissenborn et al. (eds.), 191–236.
Roeper, T. & Weissenborn, J. (1990). How to make parameters work: comments on Valian. In Frazier, L. & de Villiers, J. (eds.), *Language processing and language acquisition*. Dordrecht: Kluwer. 147–162.
Roeper, T. & Williams, E. (eds.) (1987). *Parameter setting*. Dordrecht: Reidel.
Sakas, W. G. & Fodor, J. D. (2001). The Structural Triggers Learner. In Bertolo (ed.), 172–233.
Sakas, W. G. (2003). A word-order database for testing computational models of syntax acquisition. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 415–422.
Valian, V. (1990). Null subjects: a problem for parameter-setting models of language acquisition. *Cognition* **35**. 105–122.
van Kampen, J. (1997). *First steps in* wh-*movement*. Delft: Eburon.
Villavicencio, A. (2001). *The acquisition of a Unification-Based Generalised Categorial Grammar*. Ph.D. dissertation, University of Cambridge.
Weir, R. H. (1962). *Language in the crib*. The Hague: Mouton.
Weissenborn, J., Goodluck, H. & Roeper, T. (eds.) (1992). *Theoretical issues in language acquisition: continuity and change in development*. Hillsdale, NJ: Lawrence Erlbaum.
Wexler, K. (1993). The subset principle is an intensional principle. In Reuland, E. J. & Abraham, W. (eds.), *Knowledge and language: issues in representation and acquisition*. Dordrecht: Kluwer. 217–239.
Wexler, K. (1999). Maturation and growth of grammar. In Ritchie, W. C. & Bhatia, T. K. (eds.), *Handbook of child language acquisition*. San Diego, CA: Academic Press. 55–109.

Wexler, K. & Culicover, P. W. (1980). *Formal principles of language acquisition*. Cambridge, MA: MIT Press.
Wexler, K. & Manzini, R. (1987). Parameters and learnability in binding theory. In Roeper & Williams (eds.), 41–76.
White, L. (1989). *Universal Grammar and second language acquisition*. Amsterdam: John Benjamins.
Williams, E. (1987). Introduction. In Roeper & Williams (eds.), vii–xix.
Wu, A. (1994). *The Spell-Out parameters: a Minimalist approach to syntax*. Ph.D. dissertation, UCLA.
Yang, C. D. (2002). *Knowledge and learning in natural language*. New York: Oxford University Press.

Authors' addresses: (*Fodor*)
Ph.D. Program in Linguistics, The Graduate Center, City University of New York, 365 5th Avenue, New York, NY 10016, U.S.A.
E-mail: jfodor@gc.cuny.edu

(*Sakas*)
Department of Computer Science, Hunter College, North 1008, City University of New York, 695 Park Avenue, New York, NY 10021, U.S.A.
E-mail: sakas@hunter.cuny.edu