Formal Language Theory: Refining the Chomsky Hierarchy

Gerhard Jäger & James Rogers

1 Introduction

The field of formal language theory — initiated by Noam Chomsky in the 1950s, building on earlier work by Axel Thue, Alan Turing, and Emil Post — provides a measuring stick for linguistic theories that sets a minimal limit of descriptive adequacy. Chomsky suggested a series of massive simplifications and abstractions to the empirical domain of natural language. (In particular, this approach ignores meaning entirely. Also, all issues regarding the usage of expressions, like their frequency, context dependence, and processing complexity, are left out of consideration. Finally, it is assumed that patterns that are productive for short strings apply to strings of arbitrary length in an unrestricted way.) The immense success of this framework — influencing not just linguistics to this day, but also theoretical computer science and, more recently, molecular biology — suggests that these abstractions were well chosen, preserving essential aspects of the structure of natural languages.[1]

An *expression* in the sense of formal language theory is simply a finite string of symbols, and a *(formal) language* is a set of such strings. The theory explores the mathematical and computational properties of such sets. As a starting point, formal languages are organized into a nested hierarchy of increasing complexity.

In its classical formulation [3], this so-called *Chomsky Hierarchy* has four levels of increasing complexity: regular, context-free, context-sensitive, and computably enumerable languages. Subsequent work in formal linguistics showed that this four-fold distinction is too coarse-grained to pin down the level of complexity of natural languages along this domain. Therefore several refinements have been proposed. Of particular importance here are levels that extend the class of context-free languages — the so-called *mildly context-sensitive languages* — and ones that further delimit the regular languages — the *sub-regular hierarchy*.

In this article we will briefly recapitulate the characteristic properties of the four classical levels of the Chomsky Hierarchy and their (ir)relevance to the analysis for natural languages. We will do this in a semi-formal style that does not assume specific knowledge of discrete mathematics beyond elementary set theory. On this basis, we will explain the motivation and characteristics of the mildly context-sensitive and the sub-regular hierarchies. In this way we hope to give researchers working in Artificial Grammar Learning an iron ration of formal language theory that helps to relate experimental work to formal notions of complexity.

[1] Authoratitive textbooks on this field are [11, 30].

2 The Chomsky Hierarchy

A formal language in the sense of *Formal Language Theory* (FLT) is a set of sequences, or strings over some finite vocabulary Σ. When applied to natural languages, the vocabulary is usually identified with words, morphemes or sounds.[2] FLT is a collection of mathematical and algorithmic tools about how to define formal languages with finite means, and and how to process them computationally. It is important to bear in mind that FLT is not concerned with the meanings of strings, nor with quantitative/statistical aspects like the frequency or probability of strings. This in no way suggests that these aspects are not important for the analysis of sets of strings in the real world — this is just not what FLT traditionally is about (even though it is of course possible to extend FLT accordingly — see Section 7).

To be more specific, FLT deals with formal languages (= sets of strings) that can be defined by finite means, even if the language itself is infinite. The standard way to give such a finite description is with a grammar. Four things must be specified to define a grammar: a finite vocabulary of symbols (referred to as *terminals*) that appear in the strings of the language; a second finite vocabulary of extra symbols called *non-terminals*; a special designated non-terminal called the *start symbol*; and a finite set of rules.

From now on we will assume that when we refer to a grammar $\mathcal{G}$ we refer to a quadruple $\langle \Sigma, NT, S, R \rangle$, where Σ is the set of terminals, NT is the set of non-terminals, S is the start symbol, and R is the set of rules. Rules have the form $\alpha \rightarrow \beta$, understood as meaning "α may be replaced by β", where α and β are strings of symbols from Σ and/or NT. Application of the rule "$\alpha \rightarrow \beta$" to a string means finding a substring in it that is identical with α and replacing that substring by β, keeping the rest the same. Thus applying "$\alpha \rightarrow \beta$" to $x\alpha y$ produces $x\beta y$.

$\mathcal{G}$ will be said to *generate* a string w consisting of symbols from Σ if and only if it is possible to start with S and produce w through some finite sequence of rule applications. The sequence of modified strings that proceeds from S to w is called a *derivation* of w. The set of all strings that $\mathcal{G}$ can generate is called the *language* of $\mathcal{G}$, and is notated $L(\mathcal{G})$.

The question whether a given string w is generated by a given grammar $\mathcal{G}$ is called the *membership problem*. It is *decidable* if there is a Turing machine (or an equivalent device, i.e. a computer program running on a machine with unlimited memory and time resources) that answers this question with "yes" or "no" in finite time. A grammar $\mathcal{G}$ is called *decidable* if the membership problem is decidable for every string of terminals of that grammar. In a slight abuse of terminology, a language is called decidable if it has a decidable grammar. A class of grammars/languages is called decidable if and only if all its members are decidable.

[2]This points to another simplification that is needed when applying FLT to natural languages: In each language with productive word formation rules, the set of possible words is unbounded. Likewise, the set of morphemes is in principle unbounded if loans from other languages, acronym formation and similar processes are taken into considerations. It is commonly assumed here that the object of investigation is an idealized language that does not undergo change. When the vocabulary items are identified with words, it is tacitly taken for granted that the words of a language form a finite number of grammatical categories, and that it is thus sufficient to consider only a finite number of instances of each class.

2.1 Computably enumerable languages

The class of all languages that can be defined by some formal grammar is called *computably enumerable*. It can be shown that any kind of formal, algorithmic procedure that can be precisely defined can also be expressed by some grammar — be it the rules of chess, the derivations of logic, or the memory manipulations of a computer program. In fact, any language that can be defined by a Turing machine (or an equivalent device) is computably enumerable, and vice versa.

All computably enumerable languages are *semi-decidable*. This means that there is a Turing machine that takes a string w as input and outputs the answer "yes" if and only if w is generated by $\mathcal{G}$. If w is not generated by $\mathcal{G}$, the machine either outputs a different answer or it runs forever.

Examples of languages with this property are the set of computer programs that halt after a finite number of steps (simply compile the program into a Turing machine and let it run, and then output "yes" if the program terminates), or the set of provable statements of first order logic. (A Turing machine can systematically list all proofs of theorems one after the other; if the last line of the proof equals the string in question: output "yes"; otherwise move on to the next proof.)

2.2 Context-sensitive languages

Context-sensitive grammars[3] are those grammars where the left hand side of each rule (α) is never longer than the right hand side (β). Context-sensitive languages are then the languages that can be defined by some context-sensitive grammar. The definition of this class of grammars immediately ensures a decision procedure for the membership problem. Starting from a string in question w, there are finitely many ways in which rules can be applied backward to it. None of the resulting strings is longer than w. Repeating this procedure either leads to shorter strings or to a loop that need not be further considered. In this way, it can be decided in finite time whether w is derivable from S.

Even though the question whether or not a given string w is generated by a given context-sensitive grammar $\mathcal{G}$ is in principle decidable, computing this answer may be so complex algorithmically that it is, for practical purposes, intractable.[4]

It should be noted that there are decidable languages that are not context-sensitive (even though they don't have any practical relevance in connection with natural languages).

Examples of context-sensitive languages (that are not context-free) are (we follow the common notation where x^i denotes a consecutive string of symbols that contains exactly i repetitions of the string x):

- the set of all prime numbers (where each number n is represented by a string of length n),
- the set of all square numbers,

[3]The term "context-sensitive" has only historical significance. It has noting to do with context-dependency in a non-technical sense in any way. The same applies to the term "context-free".

[4]In the terminology of computational complexity theory, the problem is PSPACE hard.

- the *copy language*, i.e. the set of all strings over Σ that consist of two identical halfs,
- $a^n b^m c^n d^m$,
- $a^n b^n c^n$, and
- $a^n b^n c^n e^n f^n$

2.3 Context-free languages

In a context-free grammar, all rules take the form

$$A \to \beta,$$

where A is a single non-terminal symbol and β is a string of symbols.[5] Context-free languages are those languages that can be defined by a context-free grammar.

Here the non-terminals can be interpreted as names of syntactic categories, and the arrow "$\to$" can be interpreted as "consists of". Therefore the derivation of a string x in such a grammar implicitly imposes a hierarchical structure of x into ever larger sub-phrases. For this reason, context-free grammars/languages are sometimes referred to as phrase structure grammars/languages, and it is assumed that such languages have an intrinsic hierarchical structure.

As hierarchical structure is inherent in many sequential phenomena in biology and culture — from problem solving to musical structure —, context-free grammars are a very versatile analytical tool.

It is important to keep in mind though that a context-free language (i.e. a set of strings) does not automatically come equipped with an intrinsic hierarchical structure. There may be several grammars for the same language that impose entirely different phrase structures.

This point can be illustrated with the language $(ab)^n(cd)^n$. A simple grammar for it has only two rules:

- $S \to abScd$,
- $S \to abcd$.

The derivation for the string *abababcdcdcd* can succinctly be represented by the *phrase structure tree* given in Figure 1. In such a tree diagram, each local tree (i.e. each node together with the nodes below it that are connected to it by a direct line) represents one rule application, with the node on top being the left-hand side and the nodes on the bottom the right-hand side. The sequence that is derived can be read off the *leaves* (the nodes from which no line extends downward) of the tree.

The same language can also be described by a somewhat more complex grammar, using the rules:

- $S \to aTd$,

[5] In context-free grammars, the right hand side of a rule may be the empty string, while in context-sensitive grammars this is not licit. Therefore, strictly speaking, not every context-free grammar is context-sensitive. This is a minor technical point though that can be ignored in the present context.

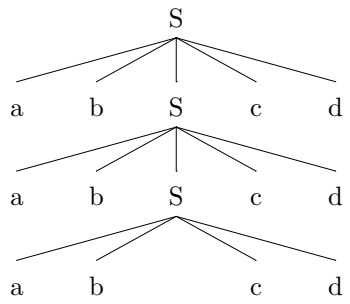


Figure 1: Phrase structure tree

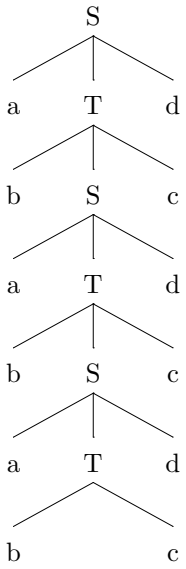


Figure 2: Different phrase structure tree for the same string

- $T \to bSc$,
- $T \to bc$.

According to this grammar, the phrase structure tree for *abababcdcdcd* comes out as given in Figure 2.

So both grammars impose a hierarchical structure on the string in question, but these structures differ considerably. It is thus important to keep in mind that phrase structures are tied to particular grammars and need not be intrinsic to the language as such.

Natural languages often provide clues about the hierarchical structure of their sentences beyond the plain linear structure. (Intonation, semantic coherence, morphological agreement and relative syntactic independence are frequently used criteria for a sub-string to be considered a coherent hierarchical unit.) Therefore most linguists require a grammar not just to generate the cor-

context-free languages	*non-context-free languages*
mirror language (i.e. the set of strings xy over a given Σ such that y is the mirror image of x)	*copy language* (i.e. the set of strings xx over a given Σ such that x is an arbitrary string of symbols from Σ)
palindrome language (i.e. the set of strings x that are identical to their mirror image)	
$a^n b^n$	$a^n b^n c^n$
$a^n b^m c^m d^n$	$a^n b^m c^n d^m$
well-formed programs of Python (or any other high-level programming language)	
Dyck language (the set of well-nested parentheses)	
well-formed arithmetic expression	

Table 1: Context-free and non-context-free languages

rect set of strings to be adequate. Rather, it must also assign a plausible phrase structure.

The membership problem for context-free languages is solvable in cubic time, i.e. the maximum time that is needed to decide whether a given string x belongs to $L(\mathcal{G})$ for some context-free grammar $\mathcal{G}$ grows with the third power of the length of x. This means that there are efficient algorithms to solve this problem.

Examples of (non-regular) context-free languages are given in the left column of Table 1. Where appropriate, a minimally differing example for a non-context-free language (that are all context-sensitive) are given in the right column for contrast.

2.4 Regular languages

Regular languages are those languages that are defined by regular grammars. In such a grammar, all rules take one of the following two forms:

$$\begin{aligned} A &\to a, \\ A &\to aB. \end{aligned}$$

Here A and B stand for non-terminal symbols and a for a terminal symbol.[6]

[6]Equivalently, we may demand that the rules take the form "$A \to a$" or "$A \to Ba$", with the non-terminal, if present, preceding the terminal. It is crucial though that within a given grammar, all rules start with a terminal on the right-hand side, or all rules end with a terminal.

regular languages	*non-regular languages*
$a^n b^m$	$a^n b^n$
the set of strings x such that the number of as in x is a multiple of 4	the set of strings x such that the number of as and the number of bs in x are equal
the set of natural numbers that leave a remainder of 3 when divided by 5	

Table 2: Regular and non-regular languages

As regular grammars are also context-free, the non-terminals can be seen as category symbols and the arrow as "consists of". According to another natural interpretation, non-terminals are names of the *states of an automaton*. The arrow "$\rightarrow$" symbolises possible state transitions, and the terminal on the right hand side is a symbol that is emitted as a side effect of this transition. The start symbol S designates the initial state, and rules without a non-terminal on the right hand side represent transitions into the final state. As there are finitely many non-terminals, a regular grammar thus describes a *finite state automaton*. In fact, it can be shown that each finite state automaton can be transformed into one that is described by a regular grammar without altering the language that is being described. Therefore regular grammars/languages are frequently referred to as *finite state* grammars/languages.

The membership problem for regular languages can be solved in linear time, i.e. the recognition time grows at most proportionally to the length of the string in question. Regular languages can thus be processed computationally in a very efficient way.

Table 2 gives some examples of regular languages in the left column. They are contrasted to similar non-regular (context-free) languages in the right column.

As the examples illustrate, regular grammars are able to count up to a certain number. This number may be arbitrarily large, but for each regular grammar, there is an upper limit for counting. No regular grammars is able to count two sets of symbols and compare their size if this size is potentially unlimited. As a consequence, $a^n b^n$ is not regular.

The full proof of this fact goes beyond the scope of this overview article, and the interested reader is referred to the literature cited. The crucial insight underlying this proof is quite intuitive though, and we will give a brief sketch.

For each regular grammer $\mathcal{G}$, it is possible to construct an algorithm (a *finite state automaton*) that reads a string from left to right, and then outputs "yes" if the string belongs to $L(\mathcal{G})$, and "no" otherwise. At each point in time, this algorithm is in one of $k+1$ different states, where k is the number of non-terminals in $\mathcal{G}$. Suppose, for a *reductio ad absurdum*, that $L = a^n b^n$ is a regular language, and let $\mathcal{G}^*$ be a regular grammar that recognizes L and that has k^* non-terminals. Then the corresponding recognition algorithm has k^*+1 different states. Now let i be some number $> k^*+1$. According to the assumption, $a^i b^i$ belongs to $L(\mathcal{G})$. When the recognition algorithm reads in the

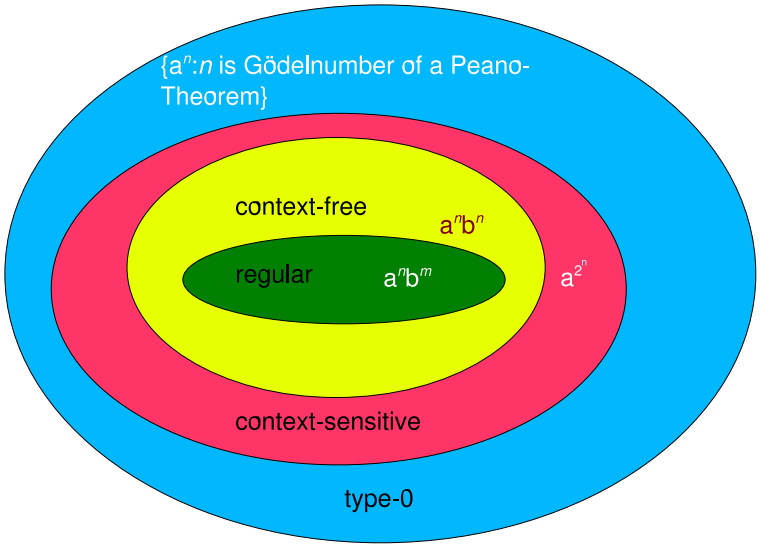


Figure 3: Chomsky Hierarchy

sequence of as at the beginning of the string, it will visit the same state for the second time after at most $k^* + 1$ steps. So a sequence of i consecutive as will be indistinguishable for the algorithm from a sequence of $i - k'$ consecutive as, for some positive $k' \leq k^* + 1$. Hence, if the algorithm accepts the string $a^i b^i$, it will also accept the string $a^{i-k'} b^i$. As this string does not belong to $a^n b^n$, the algorithm does not accept all and only the elements of $a^n b^n$, contra assumption. Therefore $a^n b^n$ cannot be a regular language.

As mentioned above, each regular language corresponds to some *finite state automaton*, i.e. an algorithm that consumes one symbol at a time and changes its state according to the symbol consumed. As the name suggests, such an automaton has finitely many states. Conversely, each finite state automaton can be transformed into a regular grammar $\mathcal{G}$ such that the automaton accepts all and only the strings in $L(\mathcal{G})$.

The other levels of the Chomsky Hierarchy likewise each correspond to a specific class of automata. Context-free grammars correspond to finite state automata that are additionally equipped with a *pushdown stack*. When reading an input symbol, such a machine can — next to changing its state — add an item on top of a stack, or remove an item from the top of the stack.

Context-sensitive grammars correspond to *linearly bounded automata*. These are essentially Turing machines, i.e. finite state automata with a memory tape that can perform arbitrary operations (writing and erasing symbols on the tape and moving the tape in either direction) during state transitions. The length of the available tape is not infinite though but bounded by a number that is a linear function of the length of the input string.

Finally, Type-0 grammars correspond to unrestricted Turing machines.

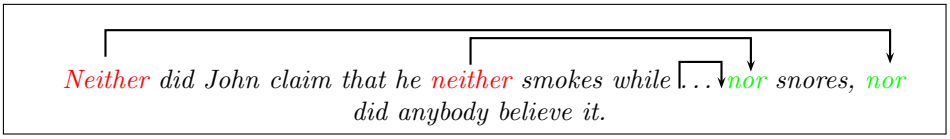


Figure 4: Nested dependencies in English

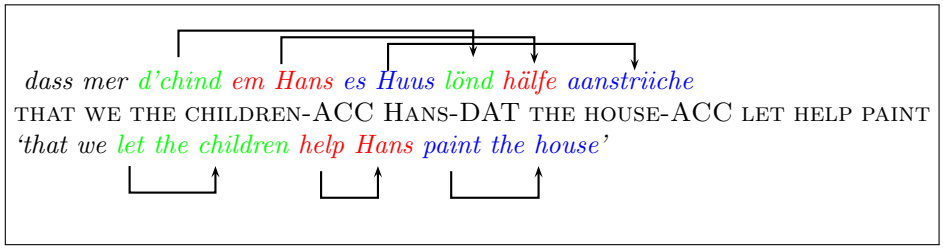


Figure 5: Cross-serial dependencies in Swiss German

3 Where are natural languages located?

The issue where natural languages are located within this hierarchy has been an open problem over decades. Chomksky [4] pointed out already in the 1950s that English is not a regular language, and this argument probably carries over to all other natural languages. The crucial insight here is that English has *centre embedding* constructions. These are constructions involving two dependent elements a and b that are not adjacent, and that may contain another instance of the same construction between the two parts. An example are *neither-nor* constructions, as illustrated in Figure 4. The pair-wise dependencies between *neither* and *nor* are nested. As far as the grammar of English goes, there is no fixed upper bound on the number of levels of embedding.[7] Consequently, English grammar allows for a potentially *unlimited number* of nested dependencies of *unlimited size*. Regular grammars are unable to recognize this kind of unlimited dependencies because this involves counting and comparing. As mentioned at the end of the previous section, regular languages cannot do this.

The issue whether all natural languages are context-free proved to be more tricky.[8] It was finally settled only in the mid-1980s, independently by the scholars Riny Huybregt ([12]), Stuart Shieber ([29]), and Christopher Culy ([7]). Huybregts and Shieber use essentially the same argument. They notice that the dependencies between verbs and their objects in Swiss German are unbounded in length. However, they are not nested, but rather interleaved so that they cross each other. An example (taken from [29]) is given in Figure 5.

Here the first in a series of three article-noun phrases (*d'chind* 'the child') is the object of the first verb, *lönd* 'let' (*lönd* requires its object to be in accusative case and *d'chind* is in accusative); the second article-noun phrase (*em*

[7]Note that here one of the idealizations mentioned above come into play here: It is taken for granted that a productive pattern — forming a *neither-nor* construction out of two grammatical sentences — can be applied to arbitrarily large sentences to form an even larger sentence.

[8]Here we strictly refer to the problem whether the set of strings of grammatical English sentences is a context-free language, disregarding all further criteria for the linguistic adequacy of a grammatical description.

Hans, 'Hans', carrying dative case) is the object of the second verb (*hälfe* 'help', which requires its object to be in dative case) and the third article-noun phrase (*es Huus* 'the house', accusative case) is the object of the third verb (*aanstriiche* 'paint', which requires an accusative object). In English, as shown in the glosses, each verb is directly adjacent to its object, which could be captured even by a regular grammar. Swiss German, however, has *crossing dependencies* between objects and verbs, and the number of these interlocked dependencies is potentially unbounded. Context-free grammars can only handle an unbounded number of interlocked dependencies *if they are nested*. Therefore Swiss-German cannot be context-free. Culy makes a case that the rules of word formation in the West-African language Bambara conspire to create unbounded crossing dependencies as well, thus establishing the non-context-freeness of this language of well-formed words.

Simple toy languages displaying the same structural properties are the copy language — where each grammatical string has the form ww for some arbitrary string w, and this creates dependencies the corresponding symbols in the first and the second half of the string — and $a^n b^m c^n d^m$, where the dependencies between the as and the cs include an unlimited number of open dependencies reaching from the bs to the ds. Therefore both languages are not context-free.

4 Mildly context-sensitive languages

After this brief recapitulation of the "classical" Chomsky Hierarchy, the rest of the paper will review two extensions that have proven useful in linguistics and cognitive science. The first one — dealt with in this section — considers levels between context-free and context-sensitive languages; so-called *mildly context-sensitive* languages. The following section is devoted to the *subregular hierarchy*, a collection of language classes that are strictly included in the regular languages.

Since the 1980s, several attempts have been made to design grammar formalisms that are more suitable for doing linguistics than the rewrite grammars from the Chomsky Hierarchy, while at the same time approximating the computational tractability of context-free grammars. The most prominent examples are Aravind Joshi's *Tree Adjoining Grammar* (see [13]) and Mark Steedman's *Combinatory Categorial Grammar* ([1, 33]). In 1991, [14] proved that four of these formalisms (the two already mentioned ones, plus Gerald Gazdar's [8] *Linear Indexed Grammars* and Carl Pollard's [20] *Head Grammars*) are equivalent, i.e. they describe the same class of languages. A series of related attempts to further extend the empirical coverage of such formalisms and to gain a deeper understanding of their mathematical properties converged to another class of mutually equivalent formalisms (including David Weir's [35] *Linear Context-Free Rewrite Systems* and *Set-Local Multi-Component TAGs*, and Ed Stabler's [31] formalisation of Noam Chomsky's [5] *Minimalist Grammars*) that describe an even larger class of formal languages. As there are no common terms for these classes, we will refer to the smaller one as TAG-languages (TAG abbreviating *Tree Adjoining Grammar*) and the larger one MG-languages (MG abbreviating *Minimalist Grammar*).

The membership problem for TAG languages is $\mathcal{O}(n^6)$, i.e. the time that the algorithm takes grows with the 6th power of the length of the string in question. Non-context free languages that belong to the TAG languages are for instance

- $a^n b^m c^n d^m$,
- the copy language,
- $a^n b^n c^n$, and
- $a^n b^n c^n d^n$.

The descriptive power of TAG languages is sufficient to capture the kind of crossing dependencies that are observed in Swiss German and Bambara.[9]

Minimalist Grammars (and equivalent formalisms) are still more powerful than that. While TAG languages may only contain up to four different types of interlocked unlimited (crossing or nesting) dependencies, there is no such upper bound for MG languages. To be more precise, each MG language has a finite upper bound for the number of different types of dependencies, but within the class of MG languages, this bound may be arbitrarily large. This leads to a higher computational complexity of the membership problem. It is still always polynomial, but the highest exponent of the term may be arbitrarily large.

Becker et al. [2] argue that this added complexity is actually needed to capture all aspects of word order variation in German.

Non-TAG languages that are included in the MG languages are for instance

- $a_1^n \dots a_k^n$ for arbitrary k, and
- w^k for arbitrary k, i.e. the k-copy language for any k.

Aravind Joshi [13] described a list of properties that an extension of the context-free languages should have if it is to be of practical use for linguistics:

- It contains all context-free languages.
- It can describe a limited number of types of cross-serial dependencies.
- Its membership problem has polynomial complexity.
- All languages in it have *constant growth property.*

With regard to the last property, let L be some formal language, and let $l_1, l_2, l_3, \dots$ be the strings in L, ordered according to length. L has the *constant growth property* if there is an upper limit for the difference in length between two consecutive elements of this list. The motivation for this postulate is that in each natural language, each sentence can be extended to another grammatical sentence by adding a single word (like an adjective or an adverb) or another short conjunct. Hence there cannot be arbitrarily large gaps in the list of possible lengths the sentences of a language can have.

This postulate excludes many context-sensitive languages, like the set of square numbers, the set of prime numbers, or the set of powers of 2 etc.

Joshi calls classes of languages with these properties *mildly context-sensitive* because they extend the context-free languages, but only slightly, preserving many of the "nice" features of the context-free languages. Both TAG languages and MG languages are mildly context-sensitive classes in this sense.

The refinement of the Chomsky Hierarchy that emerges from this line of research is displayed in Figure 6.

[9] A thorough discussion of types of dependencies in natural languages and mild context-sensitivity can be found in [32].

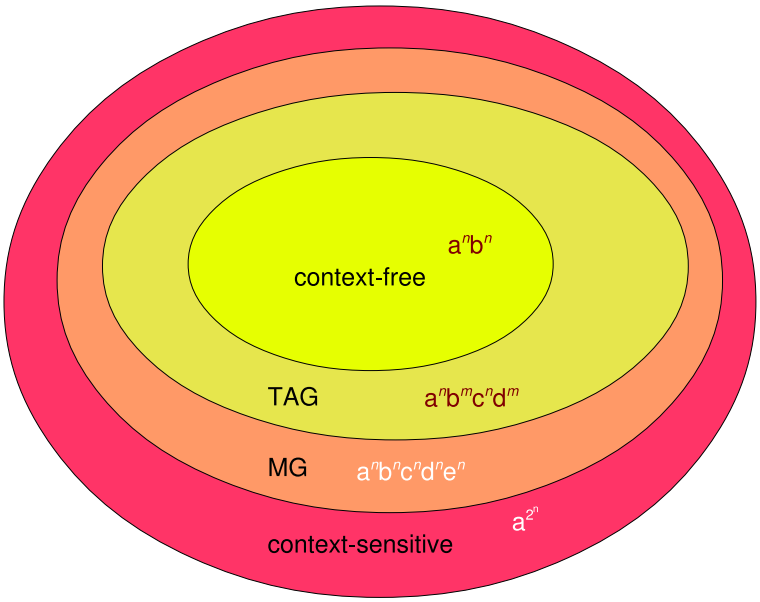


Figure 6: The mildly context-sensitive sub-hierarchy

It should be noted that Michaelis and Kracht [18] present an argument that Old Georgian is not an MG language. This conclusion only follows though if a certain pattern of complex case marking of this language is applicable recursively without limit. Of course this issue cannot be settled for a dead language, and so far the investigation of living languages with similar properties remained inconclusive. Most experts therefore assume at this time that all natural languages are MG languages.

5 Cognitive Complexity

Classes of the Chomsky Hierarchy provide a measure of the complexity of patterns based on the structure of the mechanisms (grammars, automata) that can distinguish them. But, as we observed in Section 2.3 these mechanisms make judgements about strings in terms of specific analyses of their components. When dealing with an unknown mechanism, such as a cognitive mechanism of an experimental subject, we know nothing about the analyses they employ in making their judgements, we know only that they can or cannot make these judgements about strings correctly.

The question for Artificial Grammar Learning, then, is what characteristics of the formal mechanisms are shared by the physical mechanisms employed by an organism when it is learning a pattern. What valid inferences may one make about the nature of an unknown mechanism that can distinguish the same sorts of patterns?

Here the grammar- and automata-theoretic characterisations of the Chomsky Hierarchy are much less useful. As we saw in Sections 2.4 and 4, mechanisms with widely differing natures often turn out to be equivalent in the sense that

they are capable of describing exactly the same class of languages[10]

Mechanisms that can recognise arbitrary context-free languages are not limited to mechanisms that analyse the string in a way analogous to a context-free grammar. Dependency Grammars [34], for example, analyse a string in terms of a binary relation over its elements, and there is well-defined class of these grammars that can distinguish all and only the context-free languages. In learning a context-free language, it is not necessary to analyse it in terms of immediate constituency as it is formalised by context-free grammars.

Similarly, within each of the levels of the Chomsky Hierarchy there are classes of languages that do not require the full power of the grammars associated with that level. The language $a^n b^n$, for example, while properly context-free, can be recognised by a finite state automaton that is augmented with a simple counter. The languages $a^n b^m c^m d^n$ and $a^n b^n$ with explicit nested dependencies cannot. On the other hand, these can still be recognised by mechanisms that are simpler than those that are required to recognise context-free languages in general.

So what can one say about a mechanism that can learn a properly context-free pattern? For one thing, it is not finite-state. That is, there is no bound, independent of the length of the string, on the quantity of information that it must infer in making a judgement about whether the string fits the pattern. Beyond that, there is very little if anything that we can determine about the nature of that information and how it is used simply from the evidence that an organism can learn the pattern.[11]

The situation is not hopeless, however. No matter how complicated the information inferred by a mechanism in analysing a string, it must be based on recognising simple patterns that occur in the string itself. One can, for example, identify the class of patterns that can be recognised simply on the basis of the adjacent pairs of symbols that occur in the string. Any mechanism that is based, in part, on that sort of information about the string will need to at least be able to distinguish patterns of this sort.

In the next section we introduce a hierarchy of language-theoretic complexity classes that are based on this sort of distinction: what relationships between the symbols in the string must a mechanism be sensitive to (to attend to) in order to distinguish patterns in that class. Since they are based solely on the relationships that are explicit in the strings themselves, these classes are fundamental: *every* mechanism that can recognise a pattern that is properly in one of these classes must *necessarily* be sensitive the kinds of relationships that characterise the class.

On the other hand, the fact that they are defined in terms of explicit relationships in the string itself also implies that they are all finite-state. But they stratify the finite-state languages in a way that provides a measure of complexity that is independent of the details that may vary between mechanisms that can recognise a given pattern, one that does not depend on *a priori* assumptions about the nature of the mechanism under study. Because this is a

[10] Even more strongly, it is generally possible to convert descriptions from one class of models to descriptions in an equivalent class fully automatically.

[11] This does not imply that mechanisms that are physically finitely bounded—the brain of an organism, for example—are restricted to recognising only finite-state languages. The organism may well employ a mechanism that, in general, requires unbounded memory which would simply fail when it encounters a string that is too complex, if it ever did encounter such a string.

notion of complexity that is necessarily relevant to cognitive mechanisms, and because the relative complexity of patterns is invariant over the range of equivalent mechanisms (a property not shared by measures like, for example, minimum description length) it provides a useful notion of Cognitive Complexity.

This is a notion of complexity that is particularly useful for AGL: the patterns are relatively simple, and therefore relatively practical to test, and they provide information about the capabilities of the organisms that is relevant, regardless of what additional capabilities it may have that enable it to learn more complicated patterns.

There are analogous hierarchies of classes that are based on relationships that are explicit in trees and in more complicated tree-like structures that stratify the context-free languages and a range of mildly context-sensitive languages.[24] These, do, however, apply only to mechanisms that analyse strings in terms of tree-like partial orders.

In the following section we survey a hierarchy of complexity classes that is based on adjacency within a string, the so-called Local Hierarchy. [17] There is a parallel hierarchy that is based on precedence (over arbitrary distances) that distinguishes long distance relationships within the string, including many that are relevant to a broad range of aspects of human languages—including some, but clearly not all, long distance relationships in syntax. More details of this hierarchy can be found in [25].

6 Subregular Languages

A subregular language is a set of strings that can be described without employing the full power of Finite-State Automata (FSA). Perhaps a better way of thinking about this is that the patterns that distinguish the strings that are included in the language from those that are not can be identified by mechanisms that are simpler than FSAs, often much simpler.

Many aspects of human language are manifestly subregular, including most "local" dependencies and many "non-local" dependencies as well. While these phenomena have usually been studied as regular languages, there are good reasons to ask just how much processing power is actually needed to recognise them. In comparative neurobiology, for example, there is no reason to expect non-human animals to share the full range of capabilities of the human language faculty. Even within human cognition, if one expects to find modularity in language processing then one may well expect to find differences in the capabilities of the cognitive mechanisms responsible for processing the various modules. Similarly, in cognitive evolution one would not generally expect the antecedents of the human language faculty to share its full range of cognitive capabilities; we expect complicated structures to emerge, in one way or another, from simpler ones.

The hierarchy of language classes we are exploring here are characterised *both* by computational mechanisms (classes of automata and grammars) and by model-theoretic characterisations: characterisations in terms of logical definability by classes of logical expressions. The computational characterisations provide us with the means of designing experiments: developing practical sets of stimuli that allow us to probe the ability of subjects to distinguish strings in a language in a given class from strings that are not in that language and which

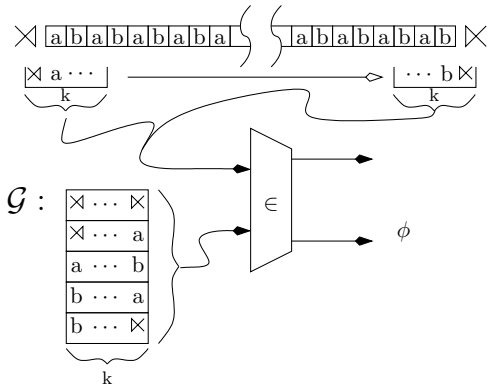


Figure 7: Scanners have a sliding widow of width k, a parameter, which moves across the string stepping one symbol at a time, picking out the k-factors of the string. For Strictly Local languages the string is accepted if and only if each of these k-factors is included in a look-up table.

suffice to resolve the boundaries of that class. The model-theoretic characterisations, because of their abstract nature, allow us to draw conclusions that are valid for all mechanisms which are capable of making those distinctions. It is the interplay between these two ways of characterising a class of languages that provides a sound foundation for designing AGL experiments and interpreting their results. *Both* types of characterisations are essential to this enterprise.

6.1 Strictly Local languages

We begin our tour of these language classes at the lowest level of complexity that is not limited to languages of finitely bounded size, patterns which depend solely on the blocks of symbols which occur consecutively in the string, with each block being considered independently of the others. Such patterns are called *Strictly Local* (SL).[12]

An SL_k definition is just a set of blocks of k adjacent symbols (called k-*factors*) drawn from the vocabulary augmented with two symbols, '$\rtimes$' and '$\ltimes$', denoting the beginning and end of the string, respectively. A string satisfies such a description if and only if every k-factor that occurs in the string is licensed by the definition. The SL_2 description $\mathcal{G}_{(AB)^n} = \{\rtimes A, AB, BA, B\ltimes\}$, for example, licenses the set of strings of the form $(AB)^n$.[13]

Abstract processing models for Local languages are called *scanners*. (See Figure 7.) For strictly k-local languages, the scanner includes a look-up table of k-factors. A string is accepted if and only if every k-factor which occurs in the string is included in the look-up table. The look-up table is formally

[12]More details on this and the other Local classes of languages can be found in [26].

[13]We use capital letters here to represent arbitrary symbols drawn from mutually distinct categories of the symbols of the vocabulary. Although none of our mechanisms involve the sort of string rewriting employed by the grammars of the first part of this paper and we distinguish no formal set of non-terminals, there is a rough parallel between this use of capital letters to represent categories of terminals and the interpretation of non-terminals as representing grammatical categories in phrase-structure grammars.

identical to an SL_k description. These automata have no internal state. Their behaviour, at each point in the computation, depends only on the symbols which fall within the widow at that point. This implies that every SL_k language will be closed under substitution of suffixes in the sense that, if the same k-factor occurs somewhere in two strings that are in the language, then the result of substituting the suffix, starting at that shared k-factor, of one for the suffix of the other must still be in the language.

Both the SL_k descriptions and the strictly k-local scanners are defined solely in terms of the length k blocks of consecutive symbols that occur in the string, taken in isolation. This has a number of implications for cognitive mechanisms that can recognise Strictly Local languages:

- Any cognitive mechanism that can distinguish member strings from non-members of an SL_k language must be sensitive, at least, to the length k blocks of consecutive symbols that occur in the presentation of the string.
- If the strings are presented as sequences of symbols in time, then this corresponds to being sensitive, at each point in the string, to the immediately prior sequence of $k-1$ symbols.
- Any cognitive mechanism that is sensitive *only* to the length k blocks of consecutive symbols in the presentation of a string will be able to recognise *only* SL_k languages.

Note that these mechanisms are *not* sensitive to the k-factors which *don't* occur in the string.

6.2 Probing the SL boundary

In order to design experiments testing an organism's ability to recognise Strictly-Local languages, one needs a way of generating sets of stimuli that sample languages that are SL and sets that sample languages that are minimally non-SL. (We return to these issues in Section 9.) This is another place in which computational characterisations of language classes are particularly useful. The language of strings of alternating symbols (e.g., 'A's and 'B's: $(AB)^n$), for example, is SL_2. Mechanisms that are sensitive to the occurrence of length 2 blocks of consecutive symbols are capable, in principle, of distinguishing strings that fit such a constraint (e.g., $(AB)^{i+j+1}$, for some i and j) from those that do not (e.g., $(AB)^i AA(AB)^j$). The ability to do so can be tested using sets of strings that match these patterns.[14]

Conversely, the language of strings in which some symbol (e.g., 'B') is required to occur at least once is not SL_k for any k. (We refer to this language as Some-B.) Mechanisms that are sensitive only to the occurrence of fixed size blocks of consecutive symbols are incapable of distinguishing strings that meet such a constraint from those that do not. Thus these organisms would not, all other things being equal, recognise that stimuli of the form A^{i+j+1} do not belong to a language correctly generalised from sets of stimuli of the form $A^i B A^j$.

[14]The patterns are both defined in terms of the parameters i and j so that the length of the strings do not vary between them.

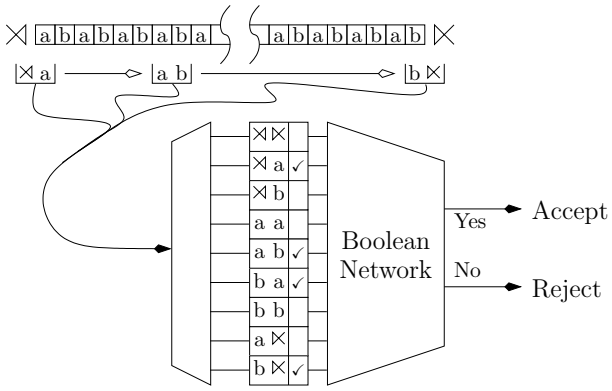


Figure 8: LT Automata keep a record of which k-factors occur in a string and feed this information into a Boolean network. The automaton accepts if, once the entire string has been scanned, the output of the network is "Yes", and rejects otherwise.

6.3 Locally k-Testable Languages

Notice that if the B were forbidden rather than required, the second pattern would be a Strictly Local (even SL_1) property. So we could define a property requiring *some* B as the *complement*—the language that contains all and only the strings that do not occur in that language—of an SL_1 language. In this case we take the complement of the set of strings in which B does not occur. If we add complement to our descriptions, it turns out that our descriptions will be able to express all Boolean operators: *conjunction* (and), *disjunction* (or), and *negation* (not) in any combination.

In order to do this, we will interpret k-factors as atomic (unanalysed) properties of strings; a string *satisfies* a k-factor if and only if that factor occurs somewhere in the string. We can then build descriptions as expressions in a *propositional logic* over these atoms. We refer to formulae in this logic as k-expressions. A k-expression defines the set of all (and only) strings that satisfy it. A language that is definable in this way is a *Locally k-Testable* (LT_k) language. The class of languages that are definable by k-expressions for any finite k is denoted LT.

By way of example, we can define the set of strings of which do not start with A and contain at least one B with the 2-expression: $(\neg \rtimes A) \wedge B$.

Note that any SL_k definition $\mathcal{G}$ can be translated into a k-expression which is the conjunction $\neg f_1 \wedge \neg f_2 \wedge \cdots$ in which the f_i are the k-factors which are *not* included $\mathcal{G}$. SL_k definable constraints are, in essence, conjunctions of negative atomic constraints and every such constraint is LT_k definable: SL is a proper subclass of LT.

A scanner for an LT_k language contains, instead of just a look-up table of k-factors, a table in which it records, for every k-factor over the vocabulary, whether or not that k-factor has occurred somewhere in the string. It then feeds this information into a combinatorial (Boolean) network which implements some k-expression. When the end of the string is reached, the automaton accepts or

rejects the string depending on the output of the network. (See Figure 8.)

Since an LT_k scanner records only which k-factors occur in the string, it has no way of distinguishing strings which are built from the same set of k-factors. Hence, a language is LT_k if and only if there is no pair of strings, each comprising the same set of k-factors, one of which is included in the language and the other excluded.

From a cognitive perspective, then:

- Any cognitive mechanism that can distinguish member strings from non-members of an LT_k language must be sensitive, at least, to the *set* of length k contiguous blocks of symbols that occur in the presentation of the string—both those that do occur and those that do not.
- If the strings are presented as sequences of symbols in time, then this corresponds to being sensitive, at each point in the string, to the set of length k blocks of symbols that occurred at any prior point.
- Any cognitive mechanism that is sensitive *only* to the occurrence or non-occurrence of length k contiguous blocks of symbols in the presentation of a string will be able to recognise *only* LT_k languages.

One of the consequences of the inability of k-expressions to distinguish strings which comprise the same set of k-factors is that LT languages cannot, in general, distinguish strings in which there is a single occurrence of some symbol from those in which there are multiple occurrences: the strings $\rtimes \underbrace{A \cdots A}_{k-1} B \underbrace{A \cdots A}_{k-1} \ltimes$ and $\rtimes \underbrace{A \cdots A}_{k-1} B \underbrace{A \cdots A}_{k-1} B \underbrace{A \cdots A}_{k-1} \ltimes$ comprise exactly the same set of k-factors. Consequently, no mechanism that is sensitive *only* to the set of fixed size blocks symbols that occur in a string will be able, in general, to distinguish strings with a single instance of a symbol from those with more than one.

6.4 Probing the LT boundary

The language of strings in which some block of k contiguous symbols is required to occur at least once (e.g., Some-B of Section 6.2, for which any $k \geq 1$ will do) is LT_k. Mechanisms which are sensitive to the set of fixed length blocks of consecutive symbols which have occurred are capable, in principle, of distinguishing strings that meet such a constraint (e.g., A^iBA^j) from those that do not (e.g., A^{i+j+1}). Again, these patterns form a basis for developing sets of stimuli that provide evidence of an organism's ability to make these distinctions.

Conversely, the language of strings in which some block of k contiguous symbols is required to occur *exactly* once (e.g., One-B, in which exactly one 'B' occurs in every string) is not LT_k for any k. Mechanisms that are sensitive only to the set of fixed length blocks of consecutive symbols which have occurred are incapable of distinguishing strings that meet such a constraint from those that do not. Thus sets of stimuli generated by the patterns A^iBA^{j+k+1} (in the set One-B) and $A^iBA^jBA^k$ (not in that set) can be used to probe whether an organism is limited to distinguishing sets of strings on this basis.

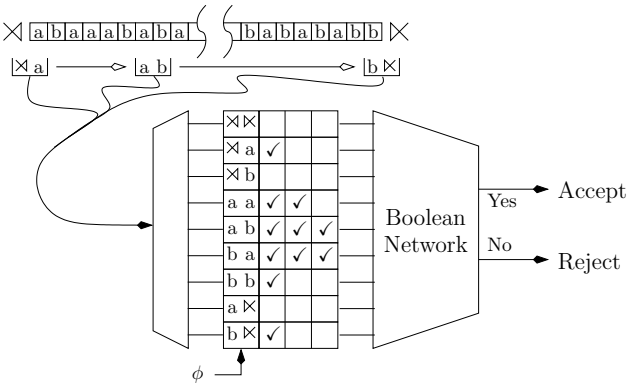


Figure 9: LTT Automata count the number of k-factors that occur in a string up to some bound and feed this information into a Boolean network. The automaton accepts if, once the entire string has been scanned, the output of the network is "Yes", and rejects otherwise.

6.5 FO(+1) definability: LTT

This weakness of LT is, simply put, an insensitivity to quantity as opposed to simple occurrence. We can overcome this by adding *quantification* to our logical descriptions, that is, by moving from Propositional logic to a First-Order logic which we call FO(+1). Logical formulae of this sort make (Boolean combinations of) assertions about which symbols occur at which positions ($\sigma(x)$, where σ is a symbol in the vocabulary and x is a variable ranging over positions in a string), about the adjacency of positions ($x \lhd y$, which asserts that the position represented by y is the successor of that represented by x) and about the identity of positions ($x \approx y$), with the positions being quantified existentially ($\exists$) or universally ($\forall$). This allows us to distinguish, for example, one occurrence of a B from another:

$$\varphi_{\text{One-}B} = (\exists x)[B(x) \wedge (\neg\exists y)[B(y) \wedge \neg x \approx y]\,]$$

This FO(+1) formula requires that there is some position in the string (call it x) at which a B occurs and there is no position (y) at which a B occurs that is distinct from that ($\neg x \approx y$).

In this example, we have defined a property expressed in terms of the occurrence of a single symbol B, but, using the successor relation, we could just as well be restricting the number of occurrences of any k-factor, for some fixed k. Moreover, using multiple levels of quantification, we can distinguish arbitrarily many distinct positions, but, since a single formula can only contain a fixed number of quantifiers, there is a fixed finite bound on the number of positions a given formula can distinguish. Hence FO(+1) formulae can, in essence, count, but only up to some fixed threshold. Note that the fixed threshold is compatible with subitization as well as actual counting.

The class of FO(+1) definable languages is is characterised by what is known as *Local Threshold Testability* (LTT). LTT automata extend LT automata by counting the number of occurrences of each k-factor, with the counters counting

up to a fixed maximum and then simply staying at that value if there are additional occurrences. (See Figure 9.)

This gives us a cognitive interpretation of LTT:

- Any cognitive mechanism that can distinguish member strings from non-members of an LTT language must be sensitive, at least, to the multiplicity of the length k blocks of symbols, for some fixed k, that occur in the presentation of the string, distinguishing multiplicities only up to some fixed threshold t.
- If the strings are presented as sequences of symbols in time, then this corresponds to being able count up to some fixed threshold.
- Any cognitive mechanism that is sensitive *only* to the multiplicity, up to some fixed threshold, (and, in particular, not to the order) of the length k blocks of symbols in the presentation of a string will be able to recognise *only* LTT languages.

6.6 Probing the LTT boundary

The language of strings in which some block of k contiguous symbols is required to occur exactly once (e.g., One-B, for which any k and $t \geq 1$ will do) is $\mathrm{LTT}_{k,t}$. Mechanisms which are sensitive to the multiplicity, up to some fixed threshold, of fixed length blocks of consecutive symbols which have occurred are capable, in principle, of distinguishing strings that meet such a constraint (e.g., A^iBA^{j+k+1}) from those that do not (e.g., $A^iBA^jBA^k$).

Conversely, the language of strings in which some block of k contiguous symbols is required to occur *prior* to the occurrence of another (e.g., No-B-after-C, in which no string has an occurrence of 'C' that precedes an occurrence of 'B', with 'A's freely distributed) is not $\mathrm{LTT}_{k,t}$ for any k or t. Mechanisms that are sensitive only to the multiplicity, up to some fixed boundary, of the occurrences of fixed length blocks of consecutive symbols are incapable of distinguishing strings that meet such a constraint from those that do not. Sets of stimuli that test this ability can be based on the patterns $A^iBA^jCA^k$, $A^iBA^jBA^k$ and $A^iCA^jCA^k$, all of which satisfy the No-B-after-C constraint, and $A^iCA^jBA^k$, which violates it.

6.7 FO(<) definability: SF

If we extend the logic of FO(+1) to express relationships in terms of precedence ($<$) as well as successor, we can define constraints in terms of both the multiplicity of factors and their order.[15] The class of FO($<$) definable languages is properly known as LTO (*Locally Testable with Order*), but this turns out to be equivalent to the better known class of *Star-Free* (SF) languages. These are the class of languages that are definable by Regular Expressions without Kleene-closure—in which the '$*$' operator does not occur—but with complement—in which the '$\overline{(\,\cdot\,)}$' operator may occur [17].

[15] The successor relationship is definable using only $<$ and quantification, so one no longer explicitly needs the successor relation. Similarly, multiplicity of factors can be defined in terms of their order, so one does not actually need to count to a threshold greater than 1.

This is, in terms of model-theoretic definability, the strongest class that is a proper subclass of the Regular languages. The Regular languages are the sets of strings that are definable using +1 (and/or <) and Monadic Second-Order quantification—quantifications over subsets of positions as well as over individual positions. It is not clear that this increase in definitional power is actually useful from a linguistic point of view. There seem to be very few, if any, natural linguistic constraints that are Regular but not Star-Free.

6.7.1 Cognitive interpretation of SF

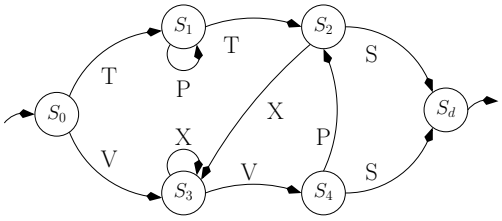
- Any cognitive mechanism that can distinguish member strings from non-members of an SF language must be sensitive, at least, to the order of the length k blocks of symbols, for some fixed k and some fixed maximum length of the sequences of blocks, that occur in the presentation of the string.
- If the strings are presented as sequences of symbols in time, then this corresponds to being sensitive to the set of sequences, up to that maximum length, of the length k blocks that have occurred at any prior point.
- Any cognitive mechanism that is sensitive *only* to the set of fixed length sequences of length k blocks of symbols in the presentation of a string will be able to recognise *only* SF languages.

7 Statistical Models of Language

Many studies of Artificial Grammar Learning have focused on statistical learning [6, 28, 27]. Language models which are based on the probability of a given symbol following another are Markov processes [15]. These can be interpreted as FSAs with transition probabilities where the underlying FSA recognises an SL_2 language. "n-gram" and "n-factor" are equivalent concepts; in general an n-gram model is a weighted version of a FSA that recognises an SL_n language; an n-gram model of a language (an $(n-1)^{\mathrm{st}}$-order Markov model) *is* a strictly n-local distribution.

Statistical models of language are not directly comparable to the sets of strings of traditional formal language theory, but there is a clear distinction between strictly local languages and n-gram models in that probabilities are not preserved under substitution of suffixes. Nevertheless, a language learner that infers probabilities of n-grams must be able to distinguish n-grams. In other words, it must attend to the n-factors of the input. Thus, the notion of cognitive complexity that we have developed here is still relevant.

Each of the levels of the hierarchy corresponds to a class of statistical distributions. The number of parameters, though, rises rapidly—the number of parameters of an LT_k distribution is exponential in k. In application, the higher complexity models are likely to be infeasible. On the other hand, there is a complexity hierarchy that parallels the local hierarchy but which is based on precedence—order of symbols independent of intervening material [25]—which also provides a basis for statistical models. The *strictly piecewise* distributions, those analogous to n-gram models, are both feasible and are capable of discriminating many long-distance dependencies [10]. The question of whether a learner



3-Factors			
$\rtimes TP$	$\rtimes TT$	$\rtimes VX$	$\rtimes VV$
TPP	TPT	TTX	TTS
TXX	TXV	$TS\ltimes$	
VXX	VXV	VVP	VVS
VPX	VPS	$VS\ltimes$	
PPP	PPT	PTX	PTS
PXX	PXV	$PS\ltimes$	
XXX	XXV	XVP	XVS

Figure 10: Reber's (1967) grammar.

can attend to subsequences (with arbitrary intervening material) in addition to or rather than substrings (factors) is significant.

8 Some Classic Artificial Grammars

Some of the earliest AGL experiments were conducted by Reber [23]. These were based on the grammar represented by the finite-state automaton in Figure 10. This automaton recognises an SL_3 language, licensed by the set of 3-factors also given in the figure—to learn a language of this form, the subject need only attend to the blocks of three consecutive symbols occurring in the strings, recognising an exception when one of the forbidden blocks occurs.

Saffran, Aslin and Newport [28] presented their test subjects with continuous streams of words in which word boundaries were indicated only by the transitional probabilities between syllables. In general, this would give an arbitrary SL_2 distribution, but the in this case the probabilities of the transitions internal to the words is 1.0 and all transitions between words are equiprobable.[16] Under these circumstances, the language is the same as that of the SL_2 automaton on which the distribution is based—i.e., this language is simply a strictly 2-local language. It should be noted, though, that from the perspective of cognitive complexity we have presented here this is a distinction without a difference. Whether the language is a non-trivial statistical model or not, to learn it the subjects need only to attend to the pairs of adjacent syllables occurring in the stimulus.

Marcus, Vijaya, Bundi Rao and Vishton [16] specifically avoided prediction by transitional probabilities by testing their subjects with strings generated according to the training pattern, but over a novel vocabulary. Gomez and Gerken [9] used a similar design. In the latter case, the grammars they used are similar to that of Reber and also license SL_3 languages. Marcus, *et al.*, limited their stimuli to exactly three syllables in order to eliminate word length as a possible cue. In general, every language of three syllable words is trivially SL_4. The focus of the experiments, though, were strings distinguished by where and if syllables were repeated (i.e. *ABA vs. AAB vs. ABB*). Languages in which no syllable is repeated are simply SL_2; those in which either the first pair of syllables, the last pair of syllables and/or the first and last syllable are repeated

[16]Technically, the final probabilities of this distribution were all 0, i.e., the distribution included no finite strings.

are SL_3. In all of these cases, the language can be distinguished by simply attending to blocks of consecutive syllables in isolation.

Finally, Saffran [27], used a language generated by a simple context-free grammar in which there is no self-embedding—no category includes a constituent which is of the same category. Hence, this language is also finite and trivially SL. Again, though, the experiments focused on the ability of the subjects to learn patterns within these finite bounds. In this case there were two types of patterns. In the first type, a word in some category occurs only in the presence of a word in another specific category. In the second type a word in some category occurs only when a word of a specific category occurs somewhere prior to it in the word. These are both non-strictly local patterns. The first is LT_1—learning this patterns requires attention to the set of syllables that occur in the word. The second is strictly 2-piecewise testable, at the lowest level of the precedence-based hierarchy. Learning it requires attention to the set of pairs of syllables in which one occurs prior to the other in the word, with arbitrary intervening material.

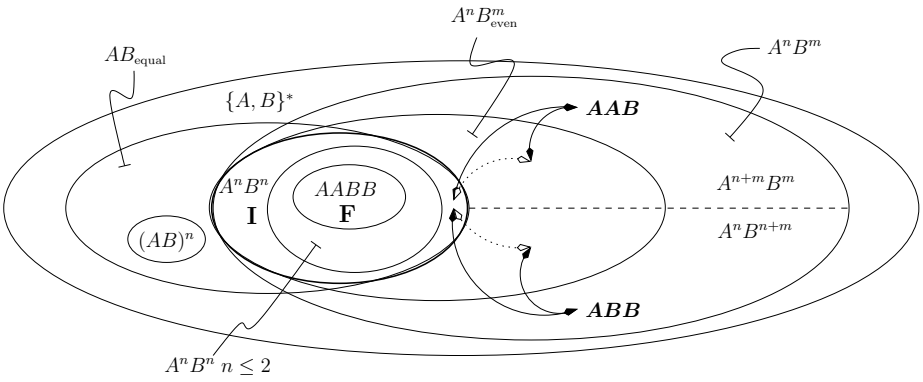
Many recent AGL experiments have employed patterns of the types $(AB)^n$ (repeated 'AB' pairs) and A^nB^n (repeated 'A's followed by exactly the same number of 'B's, sometimes with explicit pairing of 'A's and 'B's either in nested order or in "cross-serial" order: first 'A' with first 'B', etc.) As noted in Section 6.1, $(AB)^n$ is strictly 2-local, the simplest of the complexity classes we have discussed here. A^nB^n is properly context-free, with or without explicit dependencies. All automata that can recognise non-finite-state languages can be modelled as finite-state automata with some sort of additional unbounded memory (one or more counters, a stack, a tape, etc.) The question of what capabilities are required to recognise properly context-free languages, then, is a question of how that storage is organised.

As noted in Section 5, A^nB^n without explicit dependencies can be recognised by counting 'A's and 'B's (with a single counter), a simpler mechanism than that required to recognise context-free languages in general. A^nB^n with explicit nested dependencies cannot be recognised using a single counter, but it is a *linear* context-free language,[17] also simpler than the class of context-free languages in general. The question of whether there are dependencies between the 'A's and 'B's is another issue that generally depends on knowing the way that the strings are being analysed. But it is possible to make the distinction between languages that can be recognised with a single counter and those that are properly linear CFLs without appealing to explicit dependencies by using the language $A^nB^mC^mD^n$.[18] If the dependencies between the 'A's and 'B's are cross-serial then in A^nB^n is properly non-context-free. A language that makes the same distinction without explicit dependencies is $A^nB^mC^nD^m$.

The difficulty of identifying which type of structure is being used by a subject to recognise a given non-regular pattern in natural languages delayed confirmation that there were human languages that employed cross-serial dependencies for decades [22, 21, 29, 12, 7]. In AGL experiments, one has the advantage of choosing the pattern, but the disadvantage of not having *a priori* knowledge of which attributes of the symbols are being distinguished by the subjects. The fact that a particular 'A' is paired with a particular 'B' means that those instances

[17] In which only a single non-terminal occurs at any point in the derivation

[18] Note that two counters would suffice to recognise $A^nB^mC^mD^n$ but, as Minsky showed [19], two counters suffice to recognise *any* computable language.



$$I = \{A^nB^n \mid n \geq 1\} \qquad F = \{AABB\} \quad D = ?$$

Figure 11: Language inference experiments

must have a different character than other instances of the same category. Indeed, these patterns can be represented as A^nA^n with explicit dependencies of some sort between the 'A's in the first half of the string and those in the second half. Hence, most artificial grammars of this sort are actually more complicated versions of $A^nB^mC^mD^n$ or $A^nB^mC^nD^m$. There seems to be little advantage to using patterns in which the dependencies are less explicit than these.

9 Designing and Interpreting Artificial Grammar Learning Experiments

All of this gives us a foundation for exploring AGL experiments from a formal perspective. We will consider familiarisation/discrimination experiments. We will use the term *generalise* rather than "learn" or "become familiarised with" and will refer to a response that indicates that a string is recognised as an exception as *surprise*.

Let us call the set generated by the artificial grammar we are interested in **I**, the *Intended* set. The subject is exposed to some finite subset of this, which we will call **F**, the *Familiarisation* set. It then generalises **F** to some set (possibly the same set—the generalisation may be trivial). *Which* set they generalise to gives evidence of the features of **F** the subject attended to. An error-free learner would not necessarily generalise to **I**, any superset of **F** is consistent with their experience. We will assume that the set the subject generalises to is not arbitrary—it is not restricted in ways that are not exhibited by **F**—and that the inference mechanism exhibits some sort of minimality—it infers judgements about the stimuli that are not in **F** as well as those that are.[19]

We then present the subject with a set which we will call **D**, the *Discrimination* set, which includes some stimuli that are in **I** and some which are not,

[19]The assumption that the generalisation is not arbitrary implies, *inter alia*, that if it includes strings that are longer than those in **F** it will include strings of arbitrary length. This allows one to verify that a subject has not simply made some finite generalisation of the (necessarily finite) set **F**.

and observe which of these are treated by the subject as familiar and which are surprising. One can draw conclusions from these observations only to the extent that **D** distinguishes **I** from other potential generalisations. That is, **D** needs to distinguish all supersets and subsets of **I** that are consistent with (i.e., supersets of) **F**.

Figure 11 represents a situation in which we are testing the subject's ability to recognise a set that is generated by the pattern $\mathbf{I} = A^nB^n$, in which a block of 'A's is followed by block of 'B's of the same length, and we have exposed the subject to stimuli of the form $\mathbf{F} = AABB$. The feature that is of primary interest is the fact that the number of 'A's and 'B's is exactly the same, a constraint that is properly Context-Free.

The innermost circle encompasses **F**. The bold circle delimits the intended set **I**. The other circles represent sets that are consistent with **F** but which generalise on other features. For example, a subject that identifies only the fact that all of the 'A's precede all of the 'B's would generalise to the set we have called A^nB^m. This set is represented by the middle circle on the right hand side of the figure and encompasses **I**. A subject that identifies, in addition, the fact that all stimuli in **F** are of even length might generalise to the set we label $A^nB^m_{\text{even}}$. This is represented by the inner circle on the right hand side.

It is also possible that the subject has learned that there are at least as many 'A's as there are 'B's (A^nB^{n+m}) or *v.v.* These sets include **I** and that portion of A^nB^m above (resp., below) the dotted line. Alternatively, if the subject has generalised from the fact that the number of 'A's is equal to the number of 'B's but not the fact that all of the 'A's precede all of the 'B's, they might generalise to the set we label AB_{equal}. Included in this set, as a proper subset, is the language $(AB)^n$, which is not consistent with **F** but does share the feature of the 'A's and 'B's being equinumerous.

It is also possible that the subject makes the minimal generalisation that the number of 'A's is finitely bounded. The smallest of these sets consistent with **F** is the set we have labelled $A^nB^n, n \leq 2$.

These, clearly, are not all of the sets the subject might consistently infer from **F** but they are a reasonable sample of principled generalisations that we need to distinguish from **I**. Note that, in this case, the potential generalisations span the entire range of classes from SL_2 (A^nB^m), through CF (A^nB^{n+m} and AB_{equal}). If we are to distinguish, say, the boundary between Finite State and Context Free our probes will need to distinguish these. To do that, **D** must include stimuli in the symmetric differences of the potential generalisations (including **I**), that is to say stimuli that are in one or the other but not both.

For example, suppose we test the subject with strings of the form AAB. These are not in A^nB^n (properly CF) but are in the set A^nB^m (SL), so subjects that have generalised to the former will generally find them surprising while those that have generalised to the latter will not. On the other hand, they are also not in the set $A^nB^m_{\text{even}}$, which is finite state, but are in $A^{n+m}B^m$ which is properly context free. Thus these stimuli do not resolve the finite state boundary.

Suppose, then, that we test with strings of the form $AAAB$. These, again, are not in A^nB^n (CF) but are in A^nB^m (SL). But they are also not in both of $A^nB^m_{\text{even}}$ (finite state) and $A^{n+m}B^m$ (properly context free). Again, they do not resolve the finite state boundary.

One can distinguish the finite state from the non-finite state languages in

this particular set of potential generalisations if one includes strings of both of these forms in $\mathbf{D}$, but that still will not distinguish A^nB^n from A^nB^{n+m} which is presumably not significant here but may well be in testing other hypotheses. These can be distinguished by including strings of the forms that correspond to these but include more 'B's than 'A's.

None of these, though, distinguish a learner that has generalised to a finite set ($A^nB^n,\ n \leq 2$). To get evidence that the learner has done this, one needs to include strings of length greater than four.

One ends up with a discrimination set that includes at least five variants of the pattern A^nB^m for different values of n and m between two and six. This seems to be very near the boundary of practicality for most experiments involving living organisms. There are two ways that one might resolve this limitation: one can find experiments which can distinguish performance on stimuli of this size, perhaps not being able to draw any conclusions for some types of subject, or one can refine one's hypothesis so that it is practically testable. In any case, the issue is one that requires a good deal of careful analysis and it is an issue that cannot be ignored.

10 Conclusion

The notion of language theoretic complexity, both with respect to the Chomsky hierarchy and with respect to the sub-regular hierarchies is an essential tool in AGL experiments. In the design of experiments they provide a way of formulating meaningful, testable hypotheses, of identifying relevant classes of patterns, of finding minimal pairs of languages that distinguish those classes and of constructing sets of stimuli that resolve the boundaries of those languages. In the interpretation of the results of the experiments the properties of the complexity classes provide a means of identifying the pattern a subject has generalised to, the class of patterns the subject population is capable of generalising to, and ultimately, a means of identifying those features of the stimulus that the cognitive mechanisms being used are sensitive to.

In this paper we have presented a scale for informing these issues that is both finer than and broader than the finite state/context free contrast that has been the focus of much of the Artificial Grammar Learning work to date. While some of the distinctions between classes are subtle, and some of the analyses delicate, there are effective methods for distinguishing them that are generally not hard to apply and the range of characterisations of the classes provides a variety of tools that can be employed in doing so. More importantly, the capabilities distinguished by these classes are very likely to be significant in resolving the issues that much of this research is intended to explore.

Finally, fully abstract characterisations of language classes, like many of those we have presented here, provide information about characteristics of the processing mechanism that are necessarily shared by all mechanisms that are capable or recognising languages in these classes. This provides a foundation for unambiguous and strongly supported results about cognitive mechanisms for pattern recognition.

Acknowledgments

We thank William Tecumseh Fitch and the anonymous reviewers for immensely helpful comments on the draft version of this article.

References

[1] Anthony E. Ades and Mark J. Steedman. On the order of words. *Linguistics and Philosophy*, 4:517–558, 1982.

[2] Tilman Becker, Aravind Joshi, and Owen Rambow. Long-distance scrambling and tree adjoining grammars. In *Proceedings of the fifth conference on European Chapter of the Association for Computational Linguistics*, pages 21–26. Association for Computational Linguistics, 1991.

[3] Noam Chomsky. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124, 1956.

[4] Noam Chomsky. *Syntactic Structures*. Mouton, The Hague, 1957.

[5] Noam Chomsky. *The Minimalist Program*. MIT Press, Cambridge, MA, 1995.

[6] Morten H. Christiansen and Nick Chater, editors. *Connectionist Psycholinguistics*. Ablex, 2001.

[7] Christopher Culy. The complexity of the vocabulary of Bambara. *Linguistics and Philosophy*, 8(3):345–351, 1985.

[8] Gerald Gazdar. Applicability of indexed grammars to natural languages. Technical Report 85-34, Center for the Study of Language and Information, Stanford, 1988.

[9] Rebecca L. Gomez and LouAnn Gerken. Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, 70:109–135, 1999.

[10] Jeffrey Heinz and James Rogers. Estimating strictly piecewise distributions. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 886–896, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[11] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, Reading, 2000.

[12] Riny Huybregts. The weak inadequacy of context-free phrase structure grammars. In Ger Jan de Haan, Mieke Trommelen, and Wim Zonneveld, editors, *Van periferie naar kern*, pages 81–99. Foris, Dordrecht, 1984.

[13] Aravind Joshi. How much context-sensitivity is necessary for characterizing structural descriptions — tree adjoining grammars. In David Dowty, Lauri Karttunen, and Arnold Zwicky, editors, *Natural Language Processing. Theoretical, Computational and Psychological Perspectives*. Cambridge University Press, Cambridge, UK, 1985.

[14] Aravind Joshi, K. Vijay-Shanker, and David Weir. The convergence of mildly context-sensitive grammar formalisms. In Peter Sells, Stuart Shieber, and Tom Wasow, editors, *Processing of Linguistic Structure*, pages 31–81. MIT Press, Cambridge, Mass., 1991.

[15] Christopher Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[16] Gary F. Marcus, S. Vijayan, Shoba Bandi Rao, and Peter M. Vishton. Rule learning by seven-month-old infants. *Science*, 283:77–79, 1999.

[17] Robert McNaughton and Seymour A. Papert. *Counter-Free Automata*. MIT Press, 1971.

[18] Jens Michaelis and Marcus Kracht. Semilinearity as a syntactic invariant. In Christian Retoré, editor, *Logical aspects of computational linguistics*, pages 329–345. Springer, 1997.

[19] Marvin L. Minsky. Recursive unsolvability of Post's problem of "tag" and other topics in the theory of Turing machines. *Annals of Mathematics*, 74(3):437–455, 1961.

[20] Carl J. Pollard. *Generalized phrase structure grammars, head grammars and natural language*. PhD thesis, Stanford, 1984.

[21] Geoffrey K. Pullum. On two recent attempts to show that English is not a CFL. *Computational Linguistics*, 10:182–186, 1985.

[22] Geoffrey K. Pullum and Gerald Gazdar. Natural languages and context-free languages. *Linguistics and Philosophy*, 4:471–504, 1982.

[23] Arthur S. Reber. Implicit learning of artificial grammars. *Journa of Verbal Learning and Verbal Behavior*, 6:855–863, 1967.

[24] James Rogers. wMSO theories as grammar formalisms. *Theoretical Computer Science*, 293:291–320, 2003.

[25] James Rogers, Jeffrey Heinz, Gil Bailey, Matt Edlefsen, Molly Visscher, David Wellcome, and Sean Wibel. On languages piecewise testable in the strict sense. In Christian Ebert, Gerhard Jäger, and Jens Michaelis, editors, *The Mathematics of Language: Revised Selected Papers from the 10th and 11th Biennial Conference on the Mathematics of Language*, volume 6149 of *LNCS/LNAI*, pages 255–265. FoLLI/Springer, 2010.

[26] James Rogers and Geoffrey Pullum. Aural pattern recognition experiments and the subregular hierarchy. In Marcus Kracht, editor, *Proceedings of 10th Mathematics of Language Conference*, pages 1–7, 2007. University of California, Los Angeles.

[27] Jenny R. Saffran. The use of predictive dependencies in language learning. *Journal of Memory and Language*, 44:493–515, 2001.

[28] Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. Statistial learning by 8-month-old infants. *Science*, 274:1926–1928, 1996.

[29] Stuart Shieber. Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8:333–343, 1985.

[30] Michael Sipser. *Introduction to the Theory of Computation*. PWS Publishing, Boston, 1997.

[31] Edward P. Stabler. Derivational minimalism. In Christian Retoré, editor, *Logical Aspects of Computational Linguistics*, pages 68–95. Springer, 1997.

[32] Edward P. Stabler. Varieties of crossing dependencies: structure dependence and mild context sensitivity. *Cognitive Science*, 28:699–720, 2004.

[33] Mark Steedman. *The Syntactic Process*. MIT Press, Cambridge, Mass., 2000.

[34] Lucièn Tesnière. *Elèments de Syntaxe Structurale*. Klincksiek, Paris, 1959.

[35] David J. Weir. *Characterizing mildly context-sensitive grammar formalisms*. PhD thesis, University of Pennsylvania, 1988.