Opinion TRENDS in Cognitive Sciences Vol.9 No.6 June 2005

Full text provided by www.sciencedirect.com
SCIENCE DIRECT

Using mathematical models of language experimentally

Timothy J. O'Donnell[1], Marc D. Hauser[1] and W. Tecumseh Fitch[2]

[1]Primate Cognitive Neuroscience Laboratory, Harvard University, William James Hall 1052, 33 Kirkland Street, Cambridge, MA 02138, USA
[2]School of Psychology, University of St Andrews, St Andrews, Fife, KY16 9JU, UK

Understanding developmental and evolutionary aspects of the language faculty requires comparing adult languages users' abilities with those of non-verbal subjects, such as babies and non-human animals. Classically, comparative work in this area has relied on the rich theoretical frameworks developed by linguists in the generative grammar tradition. However, the great variety of generative theories and the fact that they are models of language specifically makes it difficult to know what to test in animals and children lacking the expressive abilities of normal, mature adults. We suggest that this problem can be mitigated by tapping equally rich, but more formal mathematical approaches to language.

Modern linguistics is dominated by an approach known as the generative paradigm. Generative theory is built around several core ideas about the nature of language and its study. One important cornerstone of the approach since its inception has been the way in which it envisions building theories of human language. Under the generative framework, researchers construct clear, precise models, called grammars, to describe the mature speaker's knowledge and use of language. These models are often constructed using tools from logic, mathematics and the theory of computation. The advantage of constructing such precise, and often mathematically formalized models is that the consequences of modeling decisions can be deduced directly from assumptions [1].

One of the most important questions for generative linguists and other language scientists is what biological features endow our species with its linguistic ability. Often this is expressed as the problem of determining the innate resources the child has for acquiring language [2]. It can also be reframed as an evolutionary question: Do we share some, all, or none of the key components of the language faculty with other species, and in cases where we are uniquely endowed as a species, did these capacities evolve for language in particular or for multiple domains of cognition [3,4]? Clearly such questions about language development and evolution require comparing normal adult human language abilities with the corresponding (if any) abilities in non-verbal subjects such as babies and animals.

Corresponding author: O'Donnell, T.J. (timo@wjh.harvard.edu).
Available online 5 May 2005

However, there are several challenges for the comparative experimentalist interested in turning generative theories into testable empirical predictions. First, there are a great number of generative theories of language to choose from (e.g. just in the realm of syntactic theory [5–9]). To varying degrees these theories have their own notations, their own terminology, and their own analogies with other mathematical and scientific theories. Moreover, because these theories characterize brain computations and systems of knowledge at a very high level of abstraction it is often the case that theories with radically different appearances can 'do the same work' in not-so-obvious ways (e.g. see [10]) The choice of which components of which theories to explore in non-verbal subjects is far from trivial.

Second, generative theories are built to account for the knowledge and use of normal human adult language. As such it is not always clear what they have to say about the abilities of pre- or non-linguistic subjects, such as human infants and non-human animals. It is therefore important to isolate aspects of generative theories that can be plausibly expected to show up in the non-verbal abilities of test subjects.

The subfield of mathematical linguistics offers a number of tools that allow abstract theories of language and mental computation to be compared and contrasted with one another (e.g. see [11]). We suggest that formulating comparative hypotheses with enough precision to make use of these tools can reduce some of the difficulties mentioned above. We use one set of tools from mathematical linguistics, formal language theory, to illustrate the potential power of mathematical approaches, sketching its basic concepts and discussing how they were used in one empirical example from the animal literature. We fully acknowledge that this is an area with varied opinion concerning the merits of different mathematical approaches, and the theories that back them. Brevity, however, forces us to be selective, focusing on a small corner of this potentially broad research space.

Formal language universals

One of the major goals of linguistics is to discover universal or near universal aspects of linguistic structure. Many such phenomena have been uncovered. These include things such as the way in which the presence of words in a sentence depends on the presence of other words [12], the distinction

 1364-6613/$ - see front matter © 2005 Elsevier Ltd. All rights reserved. doi:10.1016/j.tics.2005.04.011

between word positions which must be filled versus those which are optional [7], and the usefulness of constructing grammars that consist of simple structure building operations and rich sets of basic structures [5,6,8,13–17]. These also include statistical tendencies such as word order and other typological universals [18,19].

There are other universals, which are so basic that they are implicit in every linguistic theory and become most obvious when we compare language with other animal communication systems. These include the fact that language is built up from a set of reusable units, that these units combine hierarchically and recursively, and that there is systematic correspondence between how units combine and what the combination means [15,20–23].

We might expect such universals to be good targets in our search for similarities and differences between human adults and non-verbal subjects. However, even though many such universals are inherent in all linguistic theories, they are often encoded in very different ways in different theories and it is not always clear how they can be operationalized in a way that makes sense for non-verbal subjects (see Box 1).

Take, for example, the notions of constituency and dependency. These have been argued to exist in some form in every linguistic theory [12,24]. Dependency refers to the way that the presence of words in a sentence depends on other words being present in the same sentence. For instance, the definite article 'the' cannot occur in isolation without a corresponding noun. Constituency refers to the way that a sentence can be divided up into parts that can in turn be divided into parts, and so on, hierarchically, until we reach the level of words: for example, the familiar division of sentences into subject and predicate, the further division of predicates into verb and direct object, and so on. The two notions are closely related. For example, it is often the case that a constituent consists of exactly of a word and the words that depend directly on that word. However, different theories often take one or the other notion as being 'more primitive', defining one in terms of the other (see Box 1 for examples).

Given that the most basic notions of one theory may be so different from the basic notions of another, it is difficult to decide what phenomenon to look for in non-verbal subjects. For example, suppose that we find hierarchical

Box 1. The multiplicity of linguistic theories

The great number of linguistic theories reflects the degree to which even foundational issues have been debated in the field over the years. These debates go right to the core of the discipline and even concern the very nature of language as a phenomenon. Even if we restrict ourselves to generative theories, there is an enormous variety of theoretical viewpoints. For instance, Figure I shows three views of sentence (syntactic) structure chosen to illustrate the way in which different theories use different primitive notions to analyze the structure of the sentence '*John says that Fred loves Mary*'. All three structures to some degree capture the same underlying linguistic intuitions but even at a glance they can be seen to be quite different.

Limiting ourselves further to theories that have received mathematical treatment we can discern several major streams of formalization. Very early in the history of generative grammar connections were made with the theory of automata (e.g. [32]) and this has been developed extensively since. Around the same period the first definite links were made between derivation in the linguistic sense and the form of logical proofs [16,33]. More recent theoretical focus on theories as sets of constraints have led to the application of another major branch of logic, model theory, to linguistic problems [34,35]. The resource complexity of solving various linguistic problems has been studied for quite some time [29,36]. Since the early fifties there has been an interest in applying statistics and probability theory to language and over recent years this has merged to a great extent with more discrete models of language [37,38]. Approaches to language often use tools from abstract algebra [39]. Finally, the field known as 'formal language theory' is perhaps the best example of an active area of mathematical research whose origins actually lie in linguistics [27].

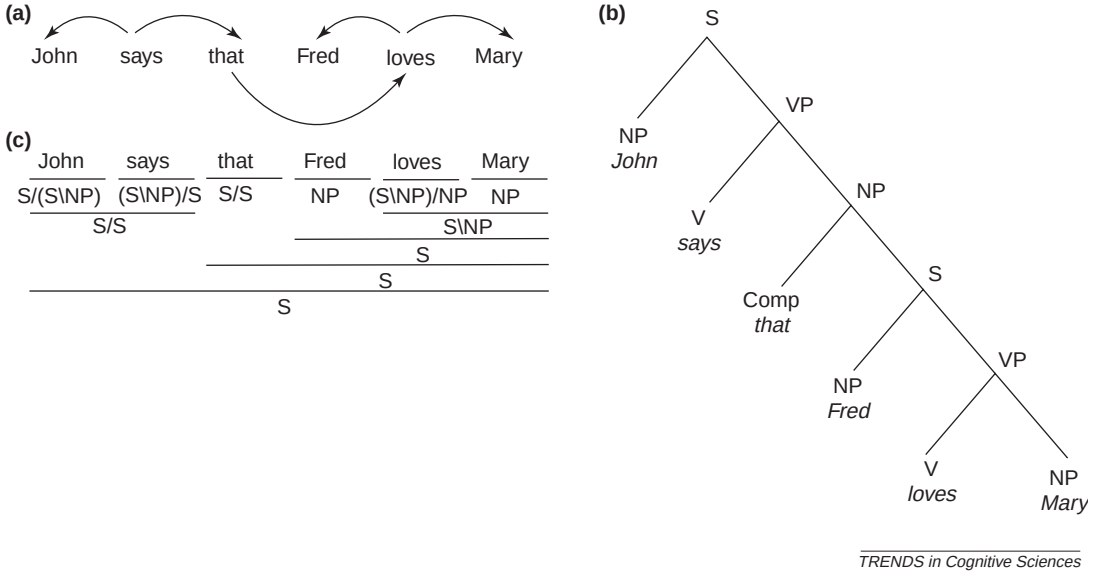

Figure I. Three ways of depicting the syntactic structure of the sentence '*John says that Fred loves Mary*'. (**a**) An analysis by means of a dependency graph, emphasizing the dependencies between the words of the sentence [12]. (**b**) A traditional phrase structure tree, which gives center stage to the linguistic notion of constituency. (**c**) A derivation based on combinatory categorial grammar that combines elements of the first two [16,31].

structure in bird song. Is this an example of analogous constituency, or is it better described by dependency, or is the question moot because they are equivalent?

Mathematical approaches to linguistics

The importance of mathematical formalization is that it allows us to think about concepts that we don't understand well, such as dependency and constituency, in terms of concepts that are well understood, such as Turing machines, sets, and functions. Usually, formalization proceeds by first defining a concept we think is important in the language of a well-understood branch of mathematics such as set theory or logic. Once we have this definition we can use mathematical principles to deduce exact consequences of our definition, and then assess how well these capture reality.

An example of this process concerns a well-known formalism known as the context-free grammars. Chomsky originally created this formalism by adapting some classic work in logic and the foundations of mathematics to model the notion of immediate constituent analysis – the idea that sentences can be hierarchically decomposed into parts until you get to the level of words [25]. One important consequence of the mathematical formalization of context-free and other types of grammars was, however, that it introduced a concept – the formal language – that allowed a wide variety of grammars to be compared with one another. It also enabled Chomsky to put an end to an important debate at the time, showing that context-free grammars can capture linguistically relevant generalizations that are simply not possible with finite-state grammars [26]. The clarity with which finite-state approaches to language were shown to be inadequate would have been impossible without mathematical formalization. This case demonstrates how mathematical approaches are capable of resolving controversies concerning the computational resources needed by adequate theories of natural language. It is our hope that such insights into the precise nature of the resources needed for language will allow us ultimately to compare the computational abilities of adult humans with possible corresponding abilities in other populations.

An example: formal language theory

Researchers have appealed to many different branches of mathematics and computer science to formalize linguistic theories, yielding a wide variety of useful tools for comparing them. As an illustrative example, we focus here on formal language theory [27] (see Box 1).

The two core concepts of formal language theory are the 'formal language' and the 'class' of formal languages. A formal language is a simple, idealized model of a set of sentences built up from some basic vocabulary. (We refer to 'sets of sentences' and 'vocabulary' as an aid to the reader in conceptualizing the elements of formal language theory, although the symbols that make up the vocabulary in a formal language can just as easily refer to phonemes, morphemes or, in fact, any level of linguistic analysis.) A class of languages is a set of formal languages satisfying some property. A property in this sense can be thought of as a statement about the languages that is true for all the languages in the class and none outside of it. In a simple toy example, we might define the property '*All sentences in the language have exactly four words*'. This gives us a class: specifically, the set of all possible sets of sentences where each sentence has exactly four words.

Usually we are concerned with classes of languages defined by reference to a type of grammatical system, for example, '*All the languages that can be described by a context-free grammar*'. The important thing about this kind of definition is that it links a group of models, context-free grammars, in this case, with the actual sets of sentences they can define. Other grammars may be very different from context-free grammars in terms of notation and primitive elements, but they too can be construed as defining sets of sentences. As a result, classes of languages provide a point of comparison for even very different models.

We prove that two kinds of grammar are different by showing that their associated classes are not identical. We demonstrate that they are the same by showing that their classes contain exactly the same languages.

In short, by giving us a way to study abstract grammatical concepts in terms of actual sets of sentences, formal language theory provides a way of comparing between linguistic theories that is less dependent on choices of notations and primitive concepts. Moreover, because in many experimental settings, such as those that use artificial languages (see Box 2), it is the actual sentences of a language that we have access to, formal language theory gives us a way to connect theory and practice.

An application of formal language theory

A recent set of experiments by Fitch and Hauser (hereafter F&H) provides an example of how formalization can be useful in comparative study of language [28]. Below we describe the steps followed in applying formal language theory to a comparative question in this study as a general example of how this might proceed.

F&H report a study looking at grammatical computation in humans and cotton-top tamarin monkeys. They use a methodology based on artificial languages in a familiarization-discrimination paradigm (see Box 2).

Step 1: Choosing a property of interest

The main tool of formal language theory is the class of languages, defined with respect to some property. In a study comparing human and non-human cognitive capacities, the goal is to choose a property that distinguishes humans from other species, in this case, humans and tamarins. F&H selected hierarchical phrase structure. In linguistic theory, hierarchy is used to describe how the elements of a sentence can be recombined to form an unbounded number of new sentences. Hierarchical structure is a ubiquitous feature of natural language syntax and F&H conjecture that it might be one of the uniquely human components of the language faculty, though not necessarily unique to language.

Step 2: Formalization

Before a property can be tested rigorously we must be able to deduce which languages are in the class in

Box 2. Experimental methods in language development and evolution

In recent years, there has been much renewed interest in experimentation with a variety of methods using artificial languages [28,40–46]. An artificial language is a set of sentences, often constructed from nonsense syllables, which has some property of interest to researchers. There have been several different experimental paradigms employed using artificial languages, but in general they use the following procedure. First subjects are exposed to some artificial language that is derived from some rule or principle. Next, subjects are given some task that requires that they discriminate between exemplars that are either consistent or inconsistent with the language they were exposed to. The use of artificial languages has the advantage of allowing the researcher to control carefully for the structure and information present in the language and focus specifically on phenomena of interest. These advantages are essential for testing human infants and non-human animals.

Formal languages can be used as models of the artificial languages in our experiments. Because formal languages can also be used as models of natural language this allows us to connect experiments with linguistic theory. Note however, that there are several challenges to this approach. Often formal languages of interest are infinite. Of course, the actual sets of sentences that we use in any given experiment are finite. In practice this means that we must test generalization cases. In these cases we test previously unheard items that require the underlying rule to judge correctly.

Figure I illustrates the difficulties inherent in testing finite sets of sentences. The finite set of sentences, 1, is contained in an infinite number of other languages. Shown here are four. Set 2 is the balanced **a** and **b** language: $\mathbf{a}^n\mathbf{b}^n$. Set 3 is the language where any number of **a**'s are followed by any number of **b**'s: $\mathbf{a}^m\mathbf{b}^w$. Set 4 might be called the 'mirror language' where each **a** in the first half of the sentence is replaced with a **b** in the second half and vice versa: $\mathbf{ww}^M$. Finally, Set 5, the set of all sequences, is also a valid hypothesis.

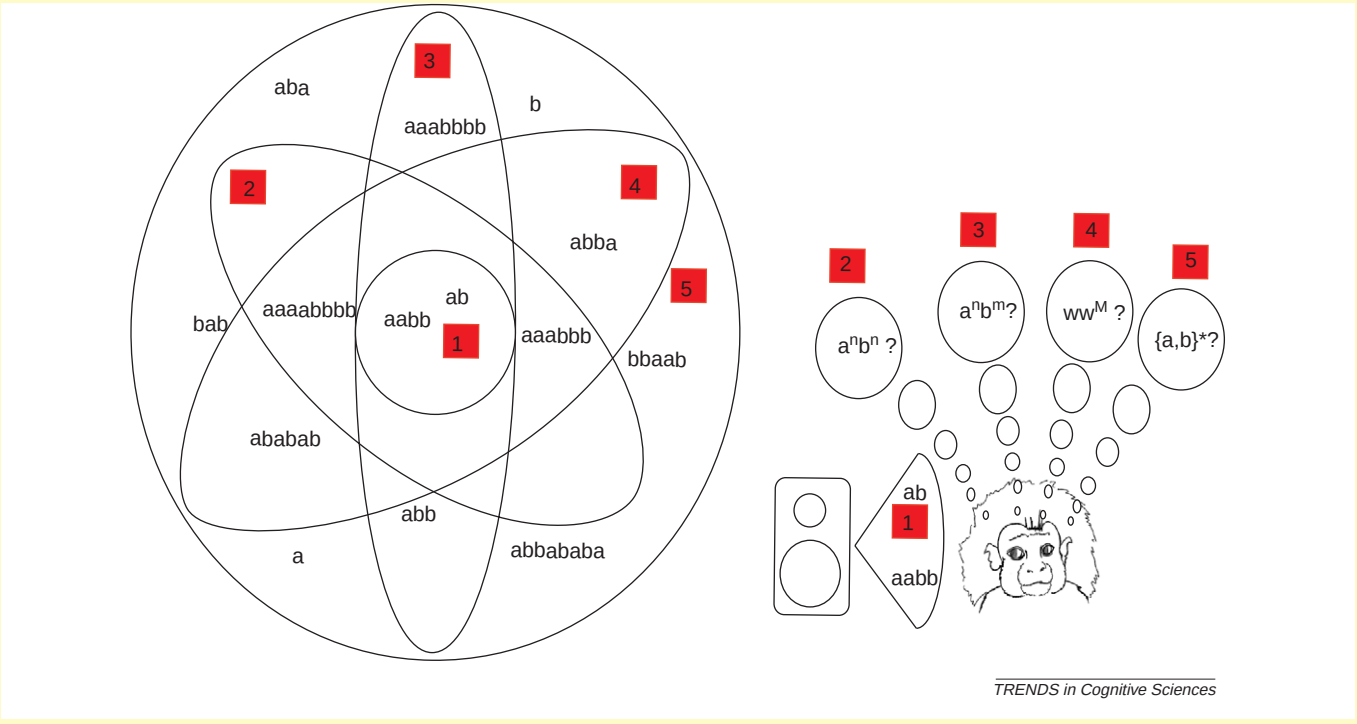

Figure I. Testing finite sets of sentences (see text for details).

question and which are not. To do this we must formalize the property. F&H do this using formal language theory. One implication of hierarchical structure is that a grammar with this property should, in a general sense, be able to generate nested components of sentences within each other without restriction. However, this implies that any procedure able to check if the subparts of a sentence are correct may have to remember an indeterminate amount of information about what it has seen so far while other parts are checked first. This property of hierarchical structures is known as 'unbounded memory'. The distinction between languages that require bounded and unbounded memory is mirrored by a distinction from formal language theory between so-called finite-state grammars and more powerful formalisms. F&H formalize structure using this distinction.

Step 3: Finding a set of sentences with the property and a set without

Following steps 1 and 2, we must choose particular languages to test. If test subjects can distinguish the two sets, then the target property describes some aspect of their computational ability. It is important here to find sets of sentences that are distinguished *only* by the property in question. To test their hypothesis, F&H selected two sets of sentences: $\mathbf{a}^n\mathbf{b}^n$ and $(\mathbf{ab})^n$. Figure 1 illustrates this choice. The first column shows several strings drawn from the first language: the 'balanced **a**'s and **b**'s' language. A variety of procedures could be used to check if a sentence is a member of this language. For example, the subject could check that each **a** matches with a **b** in a nested fashion (i), or they could check that they match in a linear fashion as in (ii). Another possibility is that they keep track of some property of the entire

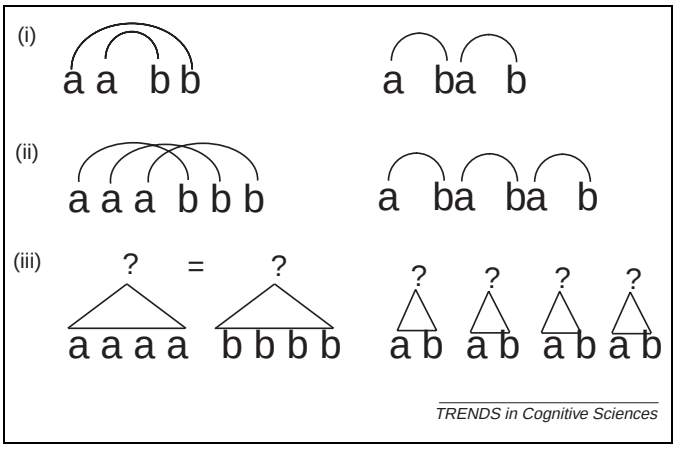

Figure 1. Strategies to check whether sentences are members of the formal languages $\mathbf{a^n b^n}$ and $\mathbf{(ab)^n}$ (see text for details).

sequence of **a**'s for example, by counting them and then seeing whether the cardinal value matches the **b**'s (iii).

The important difference between the two languages is shown in the second column: all strategies the subject uses to recognize $\mathbf{a^n b^n}$ require an unbounded amount of memory. The subject must remember a different state for every possible first half of the string, then must check that the entire second half corresponds correctly to the first. On the other hand, any strategy used for $\mathbf{(ab)^n}$ can essentially 'flush' the memory after every pair **ab** is seen, remembering only that 'all is well so far'. There is a fixed amount of memory overhead from the beginning.

Step 4: Translating the sets of sentences into exposure and test sets

The sets of sentences chosen in the previous step may or may not be directly testable in an experiment. For instance, as was the case with both test sets in F&H, they may be infinite. To get around this, F&H tested 'generalization cases' where the length, **n**, in particular test cases was different from the **n**'s used in familiarization. If the subjects extrapolate the basic pattern they will be able to apply it to longer or shorter stimuli showing that they can capture the fundamental generalization. Fair tests of the generalization phase must of course, take into account factors such as known limitations on working memory in appropriate modalities.

Conclusion

On a general level, the problem we have discussed in this article is simple: grammatical theories do not develop or evolve (in the biological senses); the neural tissue that implements particular grammars develops and the genetic programs to build that neural tissue evolve. The generative paradigm has greatly deepened our understanding of the nature of human language, but our current understanding of how the pieces of generative theories map into the circuits and networks of the brain is still very limited. Because of the variety of generative theories and their focus on describing adult language it is difficult to know what phenomena to expect to be analogous or homologous with non-linguistic abilities in non-verbal subjects. This presents a major problem for researchers interested in comparative study.

We have argued that by formalizing our theories using existing mathematical tools this problem can be mitigated. This is because theories that are formalized this way can be described using a common conceptual vocabulary with well-understood implications. Using this common conceptual basis can allow us to find precise universal properties of our linguistic systems that don't depend on notation. It may also be the case that such precisely defined properties will also be interpretable in lower level models of brain structure. For instance, we may find properties that constrain plausible neural network models.

The examples we give in this article – the combination of formal language theory, and artificial language experiments – are meant just as examples, not prescriptions. There are many other possible routes for this kind of research, and each will have its own peculiar problems. In fact, it has been observed many times that formal language theory is a rather crude tool in distinguishing systems of grammar, and there may be equal limitations with respect to empirical testing procedures [29,30]. New experimental techniques will be needed and other areas of mathematical, logical, and computational linguistics will have to be appealed to for formalizing universals. We are nonetheless optimistic that the careful application of mathematical methods in a comparative experimental setting will allow us to discover what facts about our brains make language possible, perhaps uniquely so.

Acknowledgements

This work was supported by the McDonnell 21st Century Bridging Brain, Mind, and Behavior Grant. We would like to thank Jelle Zuidema for comments.

References

1 Chomsky, N. (1957) *Syntactic Structures*, Mouton
2 Chomsky, N. (1965) *Aspects of the Theory of Syntax*, MIT Press
3 Christiansen, M. and Kirby, S., eds (2003) *Language Evolution*, Oxford University Press
4 Hauser, M. *et al*. (2002) The Language faculty: What is it, who has it, and how did it evolve? *Science* 298, 1569–1579
5 Adger, D. (2003) *Core Syntax: A Minimalist Approach*, Oxford University Press
6 Bresnan, J. (2001) *Lexical-Functional Syntax*, Blackwell
7 Haegeman, L. (1991) *Introduction to Government and Binding Theory*, Blackwell
8 Sag, I. *et al*. (2003) *Syntactic Theory: a Formal Introduction*, CSLI
9 Legendre, G. *et al*., eds (2001) *Optimality-Theoretic Syntax*, The MIT Press
10 Joshi, A. *et al*. (1991) The Convergence of Mildly Context-Sensitive Formalisms. In *Processing of Linguistic Structure* (Sells, P. *et al*., eds), pp. 31–81, MIT Press
11 Partee, B.H. *et al*. (1990) *Mathematical Methods in Linguistics*, Kluwer Academic Publishers
12 Mel'čuk, I.A. (1988) *Dependency Syntax: Theory and Practice*, State University of New York Press
13 Abeille, A. and Rambow, O., eds (2000) *Tree Adjoining Grammars: Formalisms, Linguistic Applications, and Processing*, CSLI
14 Batlin, M. and Collins, C., eds (2000) *The Handbook of Contemporary Syntactic Theory*, Blackwell
15 Jackendoff, R. (2002) *Foundations of Language*, Oxford University Press
16 Moortgat, M. (1997) Categorial type logics. In *Handbook of Logic and Language* (Bentham, J.V. and Meulen, A.T., eds), MIT Press

17 Steedman, M. (2001) *The Syntactic Process*, The MIT Press
18 Comrie, B. (1981) *Language Universals and Linguistic Typology: Syntax and Morphology*, Blackwell
19 Croft, W. (2002) *Typology and Universals*, Cambridge University Press
20 Christiansen, M.H. and Kirby, S. (2003) Language evolution: consensus and controversies. *Trends Cogn. Sci.* 7, 300–307
21 Hauser, M.D. (1997) *The Evolution of Communication*, MIT Press
22 Hockett, C.F. (1960) Logical considerations in the study of animal communication. In *Animal Sounds and Communication* (Lanyon, W.E. and Tavolga, W.N., eds), pp. 392–430, American Institute of Biological Sciences
23 Plotkin, J.B. and Nowak, M.A. (2001) Major transitions in language evolution. *Entropy* 3, 227–246
24 Miller, P.H. (2000) *Strong Generative Capacity: The Semantics of Linguistic Formalism*, CSLI
25 Bloomfield, L. (1933) *Language*, University of Chicago Press
26 Chomsky, N. (1956) Three models for the description of language. *IRE Trans. Inf. Theory* 3, 113–124
27 Hopcroft, J.E. *et al*. (2001) *Introduction to Automata Theory, Languages, and Computation*, Addison Wesley
28 Fitch, W.T. and Hauser, M.D. (2004) Computational constraints on syntactic processing in a nonhuman primate. *Science* 303, 377–380
29 Barton, G. *et al*. (1987) *Computational Complexity and Natural Language*, MIT Press
30 Conway, C.M. and Christiansen, M. (2001) Sequential learning in non-human primates. *Trends Cogn. Sci.* 5, 539–546
31 Steedman, M. (2000) *The Syntactic Process*, MIT Press
32 Chomsky, N. (1959) On certain formal properties of grammars. *Information and Control* 2, 137–167
33 Lambek, J. (1958) The mathematics of sentence structure. *Am. Math. Monogr.* 65, 154–169
34 Rogers, J. (2001) *A Descriptive Approach to Language-Theoretic Complexity*, CSLI
35 Shieber, S.M. (1992) *Constraint-Based Grammar Formalisms: Parsing and Type Inference for Natural Language and Computer Languages*, MIT Press
36 Ristad, E.S. (1993) *The Language Complexity Game*, MIT Press
37 Manning, C.D. and Schütze, H. (1999) *Foundations of Statisitical Natural Language Processing*, MIT Press
38 Pereira, F. (2000) Formal grammar and information theory: together again? *Philos. Trans. Roy. Soc.* 358, 1239–1253
39 Keenan, E. and Stabler, E. (2004) *Bare Grammar: Lectures on Linguistic Invariants*, University of Chicago Press
40 Christiansen, M. and Curtin, S. (1999) Transfer of learning: rule acquisition or statistical learning. *Trends Cogn. Sci.* 3, 289–290
41 Gomez, R.L. and Gerken, L. (2000) Infant artificial language learning and language acquisition. *Trends Cogn. Sci.* 4, 178–186
42 Onnis, L. *et al*. (2003) Reduction of uncertainty in human sequential learning: evidence from artificial language learning. In *25th Annual Conference of the Cognitive Science Scociety*, pp. 886–891, Erlbaum
43 Reber, A.S. (1967) Implicit learning of artificial grammars. *J Verbal Learn. Verbal Behav.* 6, 855–863
44 Saffran, J. (2003) Statistical language learning: mechanisms and constraints. *Curr. Dir. Psychol. Sci.* 12, 110–114
45 Saffran, J. *et al*. (1996) Statistical learning by 8-month-old Infants. *Science* 274, 1926–1928
46 Marcus, G.F. *et al*. (1999) Rule learning by seven-month-old infants. *Science* 283, 77–80

ScienceDirect collection reaches six million full-text articles

Elsevier recently announced that six million articles are now available on its premier electronic platform, ScienceDirect. This milestone in electronic scientific, technical and medical publishing means that researchers around the globe will be able to access an unsurpassed volume of information from the convenience of their desktop.

ScienceDirect's extensive and unique full-text collection covers over 1900 journals, including titles such as *The Lancet, Cell, Tetrahedron* and the full suite of *Trends* and *Current Opinion* journals. With ScienceDirect, the research process is enhanced with unsurpassed searching and linking functionality, all on a single, intuitive interface.

The rapid growth of the ScienceDirect collection is due to the integration of several prestigious publications as well as ongoing addition to the Backfiles – heritage collections in a number of disciplines. The latest step in this ambitious project to digitize all of Elsevier's journals back to volume one, issue one, is the addition of the highly cited *Cell Press* journal collection on ScienceDirect. Also available online for the first time are six *Cell* titles' long-awaited Backfiles, containing more than 12,000 articles highlighting important historic developments in the field of life sciences.

The six-millionth article loaded onto ScienceDirect entitled "Gene Switching and the Stability of Odorant Receptor Gene Choice" was authored by Benjamin M. Shykind and colleagues from the Dept. of Biochemistry and Molecular Biophysics and Howard Hughes Medical Institute, College of Physicians and Surgeons at Columbia University. The article appears in the 11 June issue of Elsevier's leading journal *Cell*, Volume 117, Issue 6, pages 801–815.

www.sciencedirect.com