



This article is part of the topic “Learning Grammatical Structures: Developmental, Cross-species and Computational Approaches,” Carel ten Cate, Clara Levelt, Judit Gervain, Chris Petkov and Willem Zuidema (Topic Editors). For a full listing of topic papers, see [http://onlinelibrary.wiley.com/journal/10.1111/\(ISSN\)1756-8765/earlyview](http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview)

On Empirical Methodology, Constraints, and Hierarchy in Artificial Grammar Learning

Willem J. M. Levelt

Max Planck Institute for Psycholinguistics

Received 14 December 2017; received in revised form 19 June 2019; accepted 19 June 2019

Abstract

This paper considers the AGL literature from a psycholinguistic perspective. It first presents a taxonomy of the experimental familiarization test procedures used, which is followed by a consideration of shortcomings and potential improvements of the empirical methodology. It then turns to reconsidering the issue of grammar learning from the point of view of acquiring constraints, instead of the traditional AGL approach in terms of acquiring sets of rewrite rules. This is, in particular, a natural way of handling long-distance dependences. The final section addresses an underdeveloped issue in the AGL literature, namely how to detect latent hierarchical structure in AGL response patterns.

Keywords: Artificial grammar learning; Text-informant presentation; Implicit priming; Constraint vs. rule learning; Long-distance dependency; Hierarchy

1. Introductory remarks

The artificial grammar learning (AGL) paradigm has brought together a diverse community of developmental linguists and psychologists, comparative biologists, cognitive neuroscientists, and others investigating the acquisition of complex sequential patterns. It has produced a rich toolkit of experimental paradigms and an equally rich variety of

Correspondence should be sent to Willem J. M. Levelt, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, the Netherlands. E-mail: pim@mpi.nl

theoretical considerations. Still, there exist remarkable gaps and unexplored possibilities in the AGL literature. The present paper considers some of them from a psycholinguistic perspective, with its long tradition of creating experimental paradigms for the study of sequential behavior and its acquisition, and with its traditional involvement with ever developing linguistic theory and its formal modeling.

We will first consider the dominant empirical paradigm in AGL, the familiarization test procedure. After presenting a more or less comprehensive taxonomy of the structures tested by means of this paradigm, I will review a number of its gaps and weaknesses and provide suggestions for improvement. One of them is a new “implicit priming” procedure, derived from a long-existing psycholinguistic experimental paradigm by the same name.

We will then turn to the core theoretical issue: What is learned in grammar learning? A consideration of the much-studied issue of long-distance dependencies invites the theoretical proposal that what is learned is a set of constraints. Constraint-based approaches are ubiquitous in linguistics; however, they are all but absent in the AGL literature, which keeps taking the traditional derivational, re-write grammar approach (explicitly or implicitly).

The final section will take up another psycholinguistic treasure, the detection of latent hierarchical structure. Fitch’s (2014) proposal that we, humans, are “dendrophiles,” loving tree structures, hierarchical patterning, has met with criticism, for instance by Honing and Zuidema (2014), who write: “it has proven very difficult to demonstrate true hierarchy at work in language and music.” It is certainly the case that the AGL literature lacks systematic procedures for detecting latent hierarchical structure in learned patterns. I propose such a procedure, at least in outline, which has been successfully applied in psycholinguistics half a century ago.

2. The two-stage familiarization test paradigm: Taxonomy of structures tested


A substantial majority of AGL studies used some version of the familiarization-test paradigm. The participant, whether human or animal, is first, in the familiarization phase presented with an experimental set of sequences. In the following test phase, the participant’s response to a set of same and/or different test sequences is measured. A large variety of sequential structures have been tested with this paradigm. Table 1 presents the types of structure tested.

The first type of structure learning I labeled “rote learning.” The subject is familiarized with a set of strings (such as *gan*, *huf*, *jom*) and subsequently tested for the memory of these strings. Saffran et al.’s (1996) original statistical learning paradigm was of this type.

The second type was correctly called “algebraic rule learning” by its inventors Marcus et al. (1999). Here, the subject is tested on recognizing adjacent or non-adjacent identity of elements in strings – such as xyx or xyx , where x and y are variables over the total vocabulary. The tests can run over analogical strings in the same vocabulary, but critically also in a different vocabulary. One frequent application of algebraic rule learning was testing the learnability of the smallest possible non-adjacency (of the xyx type).

Table 1

Types of structure tested (a, b, c for elements; x, y, z for variables; A, B for categories; m, f for male/female song motif)

Test For	Label	Structure Example	Stimulus Example	
			Exposition	Test
Same element strings	'rote'	abc	gan, huf, jom	gan, huf, jom
Analogical strings	'algebraic'	xyx		
Same vocabulary			gan, gan, jom	jom, jom, gan
Different vocabulary			gan, gan, jom	nem, nem, kov
Analogical category strings	'algebraic category'	AAB		
Same category			m_1, m_2, f_1	m_3, m_4, f_2
Different category			ga, ga, jo	
Same grammar strings	'generative'	ab^*c	gan, huf, jom	gan, huf, huf, huf, jom
		A^nB^n	ga, ga, jo, jo ($n=2$)	di, di, di, fe, fe, fe ($n=3$)

If your interest is in the human ability to acquire a grammar and possible precursors thereof, this otherwise productive rule learning paradigm does not address two central issues. The first one is "category." The syntax of natural grammars is in terms of syntactic categories, such as noun, verb, adjective, noun phrase, etcetera. It is not over the terminal vocabulary of a language, *dog, run, strong*.

Luckily, this could be solved by extending the paradigm to strings over *categories* instead of terminal elements. I called this *algebraic category*. One fine example is Seki et al. (2013) using the categories of male and female motifs in birdsong. Another one is the use of warbles vs. rattles by Comins and Gentner (2015) as motif categories. These are, doubtless, natural categories for the birds concerned. Interestingly, the birds do not acquire an xyx rule over such categories. It is, of course, possible to check whether a learned rule over categories can be transferred to a *different* category. One could for instance train subjects with syllable strings on the AAB rule and then test them on tone sequences (see Table 1).

An all-important issue is as follows: How are "abstract" syntactic categories acquired with only terminal, that is, word strings as input? Children receive inputs such as *green ball, big house, dirty mouth*; how do they derive from this an Adjective–Noun rule, which they then apply to new strings such as *hot milk*? Gerken (this issue) presents a review of this subject. Reeder et al. (2013), using an AGL paradigm, could show that adults successfully derive categories from statistical distributional input properties. But more is needed to understand the child's acquisition of linguistic categories, and there is the obvious question: Are there precursors in the animal kingdom?

The second issue can be called the "generative" one. So far, the tested rules concerned a small set of very short strings of elements, such as xyx or xyx , or of categories such as ABA or AAB . But there are no clear limits on the length of sentences that we can produce or understand. In the tradition of generative grammar this has led to mathematical models

which do not limit sentence length, such as in the Chomsky hierarchy of grammars. That these mathematical models generate infinite sets is a mathematical convenience, no more. There is no empirical fact that languages are infinite sets. Infinity of language is NOT the thing to be explained.

However, there are empirical facts that can be conveniently modeled by such grammars, for instance, the unlimited use of coordination or of modifiers, as in *John is a very very very nice fellow*, which is all easily captured by a finite state grammar which generates *John is a very* nice fellow*. Another is recursive self-embedding such as *The rat, the cat, the dog chased, killed, ate the malt*. This is nicely handled by context-free grammars. There are strong claims in the literature that recursive self-embedding is the universal core property of human language. This in spite of the fact that there are a number of live and dead languages that do not display it (Pirahã, Proto-Uralic, Dyirbal, Hixkaryana, Akkadian - cf. Pullum and Scholz (2010)) and also in spite of the fact that multiple self-embedding is quite rare in corpora of spoken language (Karlsson, 2007).

The generativity issue is addressed in the fourth type of “generative” studies addressing this issue of recursion. Two examples are given. The first one concerns strings of the type ab^*c (or category strings of the type AB^*C). An example of such a study is Chen and ten Cate (2017), to which we will return. The dominant comparison has been between strings of type A^nB^n and $(AB)^n$, introduced by Fitch and Hauser (2004). Can animals go beyond finite state grammars and learn a phrase structure grammar? Experimentally the issue is this: If you get familiarized to short strings of this type, for instance $AABB$, will you generalize to longer strings of the same type, such as $AAABBB$? Are there precursors in animals or infants of such recursive phrase structure capacity? An impressive number of experiments have been dedicated to this issue.

3. Some outstanding problems

3.1. Refining the search for learnable string patterns

The strong focus on organisms’ (in)abilities to acquire the recognition of strictly context-free stringsets, such as A^nB^n , has overshadowed the more basic problem: What are learnable patterns? Most context-free languages, including the proper subset of regular languages, are not learnable by humans. Discovering precursors of linguistic abilities in infants and non-human animals requires a broad, comparative search for what properties make stringsets learnable. Only a very small selection of stringsets has been systematically investigated in the AGL literature. Pullum and Rogers (2006) and Jäger and Rogers (2012), in particular, have pleaded for exploring the learnability of a particular hierarchy of sub-regular languages. Here, very fine contrasts can be made between levels in the hierarchy, which are potentially revealing about the organism’s cognitive abilities. At the bottom of that hierarchy are the so-called strictly local languages. The simplest cases are bigram languages, specifying which bigrams are permitted in a string. For the stringset $(AB)^n$, for instance, only AB and BA are permitted bigrams in any string. If a further

element C is added, the bigram language should specify which of the six possible bigrams AB , BA , AC , CA , BC , and CB are permitted, etc. The same can be done for tri-gram languages, and more generally for k -gram languages. The next higher level in the hierarchy consists of the “locally testable languages.” These languages consist of the strictly local languages, their unions, complements, and intersections. A very simple bigram example is the language A^nBA^m . Here AA , AB , and BA are permitted bigrams, but in addition any string should include precisely one B . This constraint, clearly, is locally testable. That also holds when the additional constraint is that B occurs *at least* once, or at least k times. At higher levels in the hierarchy, additional constraints can be built in, which allows for quite subtle tests of pattern recognition abilities. In section 4 we will return to this constraints approach.

3.2. Text versus informant presentation

The exposition procedures used in the familiarization-test paradigm are of two basic types. Half a century ago, Gold (1967) published his now classical learnability theorems. They distinguish between two modes of presentation: text and informant presentation. In text presentation the learner receives a string of legal structures/sentences, that is, positive cases. In informant presentation the learner receives both positive and negative cases, that is, also illegal structures that are marked as such. Learnability is, already for mathematical reasons, deeply different between these two presentation modes. Under Gold’s definitions, text presentation learnability is only guaranteed for finite languages, whereas informant presentation allows for learnability of finite state and context-free languages (and for “primitive-recursive” ones; cf. Levelt, 2008).

Both presentation modes have been equally used in animal AGL experiments, but the informant presentation mode is all but absent in human experimentation. A good example of the latter is the Geambasu et al. (2017) study. In this xyx/xyx -type rule learning study in adults, the two presentation modes were systematically compared and they produced dramatically different results. Informant presentation, that is, with both positive and negative feedback, not only speeded up learning, but also induced generalization of the rule to new structures. This confirms the outcome of earlier feedback studies, such as Dale and Christiansen (2004).

It is, moreover, the case that, in comparative studies, text versus informant presentation mode is usually confounded with subject population, in particular human versus animal. This should of course be avoided.

3.3. Testing procedures—off-line and on-line

The “off-line” two-phase paradigm hinges on sustained memory. However, a structure recognized by the subject during familiarization need not become a structure memorized. Memory traces may be too short-lived to be picked up in the post hoc tests. In the modern history of psycholinguistics, off-line testing has been as much as possible replaced by on-line testing in order to evade such memory problems. You want to observe

Christiansen and Chater’s (2016) “Now-or-Never” bottleneck at work while it is still on. Christiansen’s AGL-SRT task (cf. Misyak et al., 2010) has shown to be an effective on-line learning task. Gervain et al. (2012) demonstrated the power of on-line testing, getting excellent NIRS results even in sleeping newborns. It opens a window on the *dynamics* of learning and on the individual differences thereof.

3.4. An implicit priming proposal

Text versus informant presentation is a way of experimentally manipulating the participant’s attention. Another way is by *priming*. Priming became a dominant methodology in psycholinguistics some 40 years ago. Kay Bock (1986) introduced syntactic priming, which became an industry ever since (cf. Dell and Ferreira (2016)). Recently, syntactic priming studies also emerged in AGL. Fehér et al. (2016) reported an interactive AGL study, where they used syntactic priming. Kittredge and Dell (2016) used structural priming in an artificial phonetic learning study. Priming is an ideal way of affecting the learning process in precisely defined ways. It works equally well in adults, infants, and I would expect in animals. Kittredge and Dell also provide a theoretical framework which unifies priming and learning. Or, in Dell’s own terms: “learning is just priming” (personal communication).

A few decades ago Antje Meyer of my Max Planck Institute proposed the so-called implicit priming paradigm (Meyer, 1990, 1991), which meanwhile became a classical method in psycholinguistics. It is possible to apply this method *mutatis mutandis* to AG learning. The method is “on-line,” doing away with the two-phase familiarization-test paradigm and its disadvantages. I will present this alternative method by way of the example in Table 2.

The core idea of implicit priming is that you present the same item in two different environments, a homogeneous and a heterogeneous one. Homogeneous means that the item is among like items; heterogeneous, that it is among unlike items. “Like” items can mean items following the same rule or grammar. For instance, they are all of the *xyx* type

Table 2
Example of implicit priming paradigm for AGL

Comparison of homogeneous and heterogeneous blocks involving three string types: *xyx*, *xyx*, and *xyy*. The nine token strings of *Hom* are the same as the nine token strings of *Het*.

<i>Homogeneous (Hom)</i>	<i>Heterogeneous (Het)</i>
block 1: <i>xyy</i> , <i>xyy</i> , <i>xyy</i>	block 1: <i>xyy</i> , <i>xyx</i> , <i>xyy</i>
block 2: <i>xyx</i> , <i>xyx</i> , <i>xyx</i>	block 2: <i>xyx</i> , <i>xyy</i> , <i>xyx</i>
block 3: <i>xyy</i> , <i>xyy</i> , <i>xyy</i>	block 3: <i>xyy</i> , <i>xyy</i> , <i>xyx</i>

Task: (i) Read visually presented triples (such as *gan*, *gan*, *jom*) and measure speech onset RTs. Or (ii) let listen to triples and derive ERPs

Presentation: Alternate homogeneous and heterogeneous blocks, for example, *Hom1* (3 or 4 repeats), *Het1*, *Hom2*, *Het2*, *Hom3*, *Het3*

Prediction: If rule is “picked up”, responses to *the same* item in the *Hom* and *Het* conditions will be different

as in homogeneous block 1, or all of the *xyx* type as in block 2, or all of the *xyy* type as in homogeneous block 3. Now you shuffle the same items around to form three heterogeneous blocks, as on the right side in the table. Each item is now among unlike ones.

The procedure is to present alternately the homogeneous and the heterogeneous blocks to your participant, with small pauses between them. (The items in a block may be repeated any fixed number of times.) The subject must respond to each item. A human participant may for instance read the item aloud and you measure the speech onset latency. Or the subject may hear the item (such as *gan gan jom*) and repeat it and you measure the speech onset latency. Or the subject hears the item and you measure some ERP response. The prediction now is that, if a rule is “picked up” by the subject, responses to the same item will be different in the *Hom* and the *Het* conditions. Notice the power of this technique: You only make within-item comparisons, that is, the same item in a *Hom* and a *Het* condition. You also test within-participants; they are their own controls. It makes the implicit priming paradigm extremely sensitive. It is, in our experience, no exception to obtain quite significant 5–10 ms effects. You can, with enough repeats of items in a block or of blocks, also check each participant’s individual behavior. Numerous variations on this implicit priming are possible, which I confidently leave to the distinguished AGL community.

3.5. Comparing dependent variables

There is an abundance of dependent variables used in AGL experiments, among them preferential looking, grammaticality or familiarity judgments, next item prediction, eye tracking, EEG, MEG and NIRS measures, and discrimination learning. There is no reason to expect that these dependent variables measure the same thing. Almost inevitably, comparative studies use different dependent measures for the different species used. This is especially marked in the comparison of human and animal subjects. It would be no luxury to replicate classical findings in a wider range of dependent variables.

3.6. The role of semantics

The child’s acquisition of syntax is not a modular, self-contained phenomenon. It has, in particular, long been known and shown that there is substantial “semantic bootstrapping” in the acquisition of syntactic categories, such as “verb” and “noun” (Pinker, 1984). See Reeder et al. (2013) and Poletiek and Lai (2012) for reviews of this issue in the context of AGL. This approach deserves substantial elaboration. It should be entirely feasible to relate artificial grammars to presented event structures (who-does-what-to-whom, etc.) and measure their effect on grammar learning.

4. Grammar learning: Acquiring rules or constraints?

The “G” in AGL expresses the central concern: Will grammar G, however simple, be acquired from some finite set of example strings? Excluding the trivial case where the

familiarized and tested strings are identical, the “G” implicates some kind of generalization: from some finite set of strings to another finite set. The grammar G specifies the set of strings (the “language”) of which both are subsets. The AGL literature is almost exclusively based on the classical derivational approach to grammar. In this section, we will further consider the constraint-based approach to grammar, mentioned above, which may be more congenial to the study of grammar learning than grammars considered as sets of rewrite rules.

This is easily introduced by considering the issue of long-distance dependency. In the classical A^nB^n grammar learning paradigm, the long-distance dependency between the first A and the last B is the result of recursive self-embedding. The rules $S \rightarrow ASB$ and $S \rightarrow AB$ generate the self-embedding phrase structure hierarchy. However, hierarchy is not a condition for long-distance dependencies in strings.

Consider Fig. 1. It presents quite a simple finite state automaton, which accepts the regular language $\{ac^*a \cup bc^*b\}$. It is fully recursive and displays unlimited long-distance dependency: If a string begins with a , it should end with a ; if it begins with b , it should end with b . Neither hierarchy building nor recursive self-embedding is at issue here as in context-free grammars. Long-distance dependency does not at all require pushdown storage.

Still, it is an interesting type of unlimited non-adjacent dependency. Chen and ten Cate (2017) had the same insight and were the first to test zebra finches’ sensitivity to long-distance dependencies generated by this type of finite state grammar. They took care that, different from the Fig. 1 automaton, the dependency was between two non-identical elements. They tested dependencies over up to three intervening elements. Using a go/no-go familiarization procedure (i.e., informant presentation), they discovered that their finches did acquire this type of long-distance dependency.

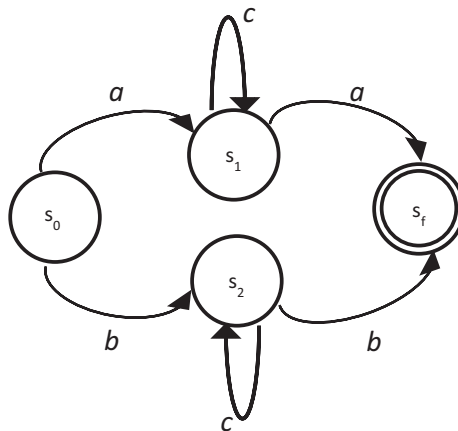


Fig. 1. Finite state automaton long-distance, generating language $\{ac^*a \cup bc^*b\}$. The language displays unlimited recursion and unlimited long-distance dependency: Any string beginning with a will also end with a ; any string beginning with b will also end with b .

This kind of dependency is quite natural in language processing, much more natural and ubiquitous than self-embedding. It works universally in agreement phenomena. When you happen to hear a singular subject, *the horse*, you expect to get a singular main verb, such as *runs*. But when the subject is plural, *the horses*, so will be the verb that is in the offing: *run*. In such cases of agreement there is no clear limit on the intervening material. You just know “some time it will come”—for the listener it is a simple expectation. That, apparently, has precursors in the animal kingdom.

A natural way of representing an organism’s awareness of dependencies, long or short distance, is by means of *constraints*. If the human or animal has captured the dependences generated by the finite state automaton in Fig. 1, it has acquired a number of constraints:

- i. Any string beginning with *a* will end with *a*
- ii. Any string beginning with *b* will end with *b*

A further constraint captures what can happen between beginning and end:

- iii. Between beginning and end there are any number of *c*’s

Constraints have always been around in linguistics. The core of government-binding theory, for instance, was a set of constraints. Optimality theory is a theory of constraint ranking. The treatment of phonotactics, in particular, is usually in terms of constraints, such as the consonantal constraints in Arabic discussed by Adriaans and Kager (2010). Constraint-based formalisms have become the natural way to handle agreement phenomena, such as number and person agreement between a subject NP and its predicate VP as in *the horses run*. The constraint on the *unification* of the phrases *the horse* and *run* is that they agree in person and number. See Jurafsky and Martin (2008), chapter 16, for a review of this constrained-based approach to unification.

This alternative constraint-based approach to grammar, the so-called Model Theoretic Syntax (MTS), has been developed by Rogers (2003) and others; see Pullum (2013) for a comparison of the two approaches. In the latter paper, Pullum considered the relevance of MTS for the handling of language acquisition:

MTS fits naturally with a very different view of first language acquisition: incremental amassing of constraints in a way that facilitates increasingly improved matching with other speakers. Notice, the constraint system acquired need only be roughly comparable to those of other speakers. No recursive specification of a target set of expressions must be attained, and there is no necessity for the internal representation of the overall effect of the assumed constraints to be similar between individuals. Humans are extraordinarily tolerant of divergence and error, and approximate similarity of observed consequences will suffice to permit conversation. (Pullum, 2013, p. 510)

What children acquire is, from this perspective, not a set of rewrite rules, but rather a set of constraints. These are acquired one by one, such increasingly matching the input they receive from other speakers. At any one stage, the child’s grammar consists of the acquired constraints, not an acquired set of rewrite rules. The more constraints they

acquire, the more “grammatical” their language becomes. Grammaticality comes in degrees.

The acquisition of constraints in humans or animals could be handled as Bayesian statistical learning. The constraints have priors, initial sensitivities in the infants or baby birds. The body of evidence, the speech spoken to children, the song listened to by birds, will increase or decrease the probability of these constraints. Lipkind et al.’s (2013) findings on the acquisition of vocalizations in Bengalese finches and in the babbling of infants can, I suppose, be put in terms of growing constraints.

In natural language, long-distance dependencies often do involve hierarchies. In the sentence, *The horse which we ride runs* the relevant agreement relation is between *the horse* and *runs*, not between *the horse* and *ride*. The constraint on their unification spans the embedded relative clause *which we ride*. Hierarchical embedding is part and parcel of natural syntax. Lashley (1951) refocused us on the hierarchical, multilevel organization of sequential behavior in both humans and animals, our task being in both cases to study the “syntax of act” (p. 188); see Levelt (2014) for the historical context in which this paper appeared. However, the search for hierarchy in the sequential behavior of animals is a challenging endeavor. The AGL literature, in particular, doesn’t provide methodologies for detecting whether hierarchical patterns have been acquired. Here again, the psycholinguistic literature may provide some help. That is the topic of the next, final section of this paper.

5. Empirical tests for hierarchical structure

The more general empirical issue is whether the human or animal subject recognizes or memorizes a particular sequence as hierarchically organized, that is, in phrases and sub-phrases. Levelt (1969, 2008) developed a procedure for testing whether such is the case. It was developed for testing syntactic hierarchies, but the algorithm involved is mathematically general, that is, content-independent. Here I will, for ease of exposition, discuss the procedure from a syntactic example. Consider the sentence *Anne buys cheap darts (abcd)*. Do we conceive of this string as hierarchically organized? Fig. 2 represents two of the five possible binary phrase structures (PS) for this sentence. PS(i) is the linguistically common one: There is the noun phrase (NP) *cheap darts*, which is embedded in the verb phrase (VP) *buys cheap darts*, which is embedded in the sentence (S) *Anne buys cheap darts*. But four other binary bracketings are possible, of which PS(ii) is an example. Is it possible to determine empirically which phrase structure is the correct one or whether *any* phrase structure is correct?

It is, but one needs empirical data of a particular kind. Needed is some measure of pairwise relatedness among the words in the sentence. Intuitively *cheap* and *darts* are stronger related than *Anne* and *cheap*. One can ask subjects to rank order such pairs in terms of relation strength, given the sentence, and they may judge $r(\textit{cheap}, \textit{darts}) > r(\textit{Anne}, \textit{cheap})$ or in short $r(c,d) > r(a,c)$. How can one theoretically relate such data to a possible phrase structure of the sentence?

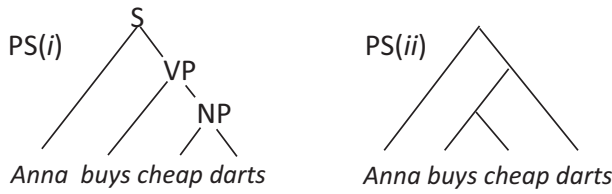


Fig. 2. Two of the five possible phrase structure (PS) trees for the sentence *Anna buys cheap darts*.

In Levelt (2008, vol. III, section 2.4.2), I proposed to define a “cohesion function” α over the nodes of phrase structures; see Fig. 3. If phrase B is embedded in phrase A , the cohesion of the embedded phrase B , $\alpha(B)$, is greater than the cohesion of the embedding phrase, $\alpha(A)$. In example PS(i), the cohesion of the most embedded phrase *cheap darts*, $\alpha(cd)$, is greater than the cohesion of *buys cheap darts*, $\alpha(bcd)$, which in turn is greater than $\alpha(abcd)$, the cohesion of the whole sentence.

Now define the “smallest common constituent” (SCC) of two words in the sentence as the smallest phrase to which they both belong. $SCC(cheap, darts)$ in PS(i) is the noun phrase *cheap darts* and $SCC(Anne, cheap)$ is the whole sentence *Anne buys cheap darts*. It is now possible to relate relation strength to cohesion as follows: for all words i, j, k, l in the sentence $r(i,j) > r(k,l) \Leftrightarrow \alpha SCC(i,j) > \alpha SCC(k,l)$. In the above example $r(c,d)$ was judged greater than $r(a,c)$. That agrees with the cohesion $\alpha(cd)$ being greater than the cohesion $\alpha(abcd)$. Although only inequalities are formulated in the if-and-only-if statement, it follows by exclusion that equal degrees of relatedness go with equal cohesion values. The relatedness $r(Anne, cheap)$, for instance, is identical to $r(Anne, darts)$ because they have the same smallest common constituent, namely the whole sentence.

If this is the way pairwise relatedness of words in the sentence is related to hierarchical phrase structure, a very interesting inequality holds:

1. *Ultrametric inequality*: $r(x,y) \geq \min(r(x,z), r(y,z))$, where x, y, z are any words in the sentence.

Levelt (2008, vol. III, 2.4.2) provides a detailed treatment. The key point is that if a string is hierarchical, the ultrametric inequality holds among its relatedness values. Also, the reverse holds: If the relatedness values among elements in a string obey the ultrametric inequality, the string is hierarchical. This was proven by Johnson (1967).

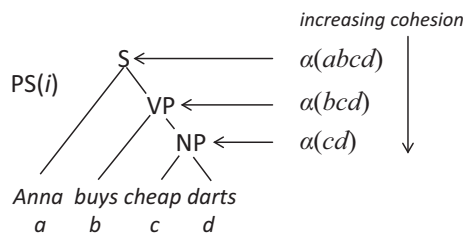


Fig. 3. Cohesion values α for phrases in phrase diagram PS(i), monotonically increasing from top to bottom.

Since Johnson's classical paper a large variety of the "hierarchical clustering algorithms" have been developed (see Blashfield and Aldenderfer (1988) and Yim and Ramdeen (2015) for reviews). They map a symmetrical relatedness matrix onto some statistically best-fitting hierarchy or branching tree. For example, if our experiment provides us with the symmetrical 4×4 relatedness matrix for the sentence *Anne buys cheap darts* and it meets the ultrametric inequality (within measurement error), Johnson's algorithm will select the best-fitting phrase diagram in Fig. 2.

The input to any such analysis is a relatedness matrix, containing a measure of relatedness for any pair of elements in the experimental string or sequence. In the human case, judgmental scaling data are easily obtained, as in the above example. But that is much harder in the animal case. Can one conceive of behavioral methods providing measures of pair-wise relation strength? For the human case, I developed such a behavioral method, which *mutatis mutandis* would be applicable in the animal case. Levelt (1970) had 120 participants listen to sentences disturbed by white noise. After each presentation, they wrote down what they had heard. For each sentence, a table of conditional probabilities $p(j/i)$ was computed, where $p(j/i)$ is the probability that word j had been correctly identified, given correct identification of word i . This was done for all i 's and j 's from the sentence. These data were highly ultrametric. Hierarchical cluster analysis of the off-diagonal submatrices for which words i precede words j revealed the hierarchical structure of major phrases in the experimental sentences. These hierarchical patterns could have been due to the subjects' incremental "chunking" of the input word string, or alternatively to their procedures of retrieving the string from memory, that is, by phrase and sub-phrase.

Could a similar behavioral measure be developed in the case of songbirds? One wonderful example in the literature comes quite close. Hultsch and Todt (2004) reported a study in which common nightingales were tutored on experimental strings of kind-specific songs. Dependent on the experimental design, a training string would contain some 20–25 songs that were all different. The duration of such songs is about 3 s. Each song consists of about eight or nine syllables, but the experimental unit was the full song. Eighteen 3- to 7-week-old male nightingales were tutored on these strings. Their first crystalized songs were recorded some 40 weeks later. It was possible to identify the training songs in these reproductions, ignoring minor variations. This allowed the authors to analyze how the reproduced strings related to the tutored strings. One experimental variable was this: Pauses in the training strings had a 4-s duration, but either two or three longer, 20-s pauses were inserted in the experimental strings. Would this invite the young nightingale to "chunk" the string in substrings, separated by the longer pauses? That is indeed what was found. The birds sometimes reproduced the whole string, more often larger or smaller "packages" of its songs. The partial reproductions were dominantly "coherent" portions from the training substrings, rarely crossing substring (20 s) boundaries. In other words, the reproductions showed a latent two-level hierarchy. The authors concluded from this experiment and further ones that the observed latent hierarchical structure results from the birds' memory storage procedures during tutoring.

The analyses involved computing the frequencies at which song i would be followed by song j in the birds' reproductions. This comes close to the analysis in Levelt (1970),

sketched above. It would indeed be possible to compute the full off-diagonal submatrix of probabilities $p(j/i)$ of a bird's producing song j , given that song i had been produced in a particular song sequence. Such a matrix can be subjected to the same type of hierarchical cluster analysis as used in Levelt (1970). The resulting clusters would reveal the birds' preferred "packages" and sub-packages. This would invite experimental designs in which the song strings presented during the phase of tutoring induce deeper hierarchies.¹ The nightingales might receive high-frequency presentations of different short 2–4 item song packages, less frequent presentations of particular 2–3 item sequences of these small song packages, and still less frequent presentations of the full concatenation of the latter larger packages. Would such a deeper hierarchy show up in the birds' first crystalized songs?

A quite different procedure for detecting hierarchy in produced sequential patterns makes use of their temporal structure (cf. Falk & Kello, 2017). It tests in essence the degree to which a temporal sequence of events follows a Poisson distribution. The more events "cluster," the larger the deviation from Poisson. This method, however, does not yield a "best-fitting" hierarchy over the sequence of events.

Note

1. I am grateful to Tecumseh Fitch for thorough discussions of potential experimental paradigms allowing for the measurement of latent hierarchical structure in birds' (or other animals') organization of sequential patterns. The reference to this work on nightingales is his.

References

- Adriaans, F., & Kaager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, *62*, 311–331.
- Blashfield, R. K., & Aldenderfer, M. S. (1988). The methods and problems of cluster analysis. In J. R. Nesselroade, & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 447–473). New York, NY: Plenum Press
- Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355–387.
- Chen, J., & ten Cate, C. (2017). Bridging the gap: Learning of acoustic nonadjacent dependencies by a songbird. *Journal of Experimental Psychology: Animal Learning and Cognition*, *43*, 295–302.
- Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, 1–72.
- Comins, J. A., & Gentner, T. Q. (2015). Pattern-induced covert category learning in songbirds. *Current Biology*, *25*, 1873–1877.
- Dale, R., & Christiansen, M. H. (2004). Active and passive statistical learning: Exploring the role of feedback in artificial grammar learning and language. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*, 262–267.
- Dell, G. S., & Ferreira, V. S. (2016). Thirty years of structural priming: An introduction to the special issue. *Journal of Memory and Language*, *91*, 1–4.

- Falk, C., & Kello, C. T. (2017). Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition*, *163*, 80–86.
- Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, *91*, 158–180.
- Fitch, W. F. (2014). Toward a computational framework for cognitive biology: unifying approaches from cognitive neuroscience and comparative cognition. *Physics of Life Reviews*, *11*, 329–364.
- Fitch, W. T., & Hauser, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science*, *303*, 377–380.
- Geambasu, A., Spierings, M. J., Levelt, C., & ten Cate, C. (2017). What is necessary to learn rules? Poster, Lorenz Workshop the Comparative Biology of Language Learning.
- Gerken, L. A. (this issue). Infant language learning in the lab.
- Gervain, J., Berent, I., & Werker, J. F. (2012). Binding at birth: The newborn brain detects identity relations and sequential position in speech. *Cognitive Neuroscience*, *24*, 564–574.
- Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*, 441–474.
- Honing, H., & Zuidema, W. (2014). Decomposing dendrophilia. Comment on "Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition" by W. Tecumseh Fitch. *Physics of Life Reviews*, *11*, 375–276.
- Hultsch, H., & Todt, D. (2004). Approaches to the mechanisms of song memorization and singing provide evidence for a procedural memory. *Anais da Academia Brasileira de Ciências*, *76*, 219–230.
- Jäger, G., & Rogers, J. (2012). Formal language theory: refining the Chomsky hierarchy. *Philosophical Transactions of the Royal Society B. Biological Sciences*, *367*, 1956–1970.
- Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*, 241–254.
- Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing* (2nd ed). New York: Prentice Hall.
- Karlsson, F. (2007). Constraints on multiple center-embedding of clauses. *Journal of Linguistics*, *43*, 365–392.
- Kittredge, A. K., & Dell, G. S. (2016). Learning to speak by listening: Transfer of phonotactics from perception to production. *Journal of Memory and Language*, *89*, 8–22.
- Lashley, K. (1951). The problem of serial order in behavior. In L. A. Jeffers (Ed.), *Cerebral mechanisms in behavior* (pp. 112–136). New York: Wiley.
- Levelt, W. J. M. (1969). The scaling of syntactic relatedness: A new method in psycholinguistic research. *Psychonomic Science*, *17*, 351–352.
- Levelt, W. J. M. (1970). Hierarchical chunking in sentence processing. *Perception & Psychophysics*, *8*, 99–103.
- Levelt, W. J. M. (2008). *Formal grammars in linguistics and psycholinguistics*. Amsterdam: John Benjamins.
- Levelt, W. J. M. (2014). *A history of psycholinguistics. The pre-Chomskyan era*. Oxford, UK: Oxford University Press.
- Lipkind, D., Marcus, G. F., Bemis, D. K., Sasahara, K., Jacoby, N., Takahasi, M., Suzuki, K., Feher, O., Ravbar, P., Okanoya, K., & Tchernichovski, O. (2013). Stepwise acquisition of vocal combinatorial capacity in songbirds and human infants. *Nature*, *498*, 104–108.
- Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*, 77–80.
- Meyer, A. S. (1990). The time course of phonological encoding in language production: The encoding of successive syllables of a word. *Journal of Memory and Language*, *29*, 524–545.
- Meyer, A. S. (1991). The time course of phonological encoding in language production: Phonological encoding inside a syllable. *Journal of Memory and Language*, *30*, 69–69.
- Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science*, *2*, 138–153.
- Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

- Poletiek, F. H., & Lai, J. (2012). How semantic biases in simple adjacencies affect learning a complex structure with non-adjacencies in AGL: A statistical account. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 367, 2046–2054.
- Pullum, G. K. (2013). The central question in comparative syntactic metatheory. *Mind & Language*, 28, 492–521.
- Pullum, G. K., & Rogers, J. (2006). *Animal pattern-learning experiments: Some mathematical background*. Boston: Radcliffe Institute for Advanced Study/Harvard.
- Pullum, G. K., & Scholz, B. C. (2010). Recursion and the infinitude claim. In H. van der Hulst (Ed.), *Recursion and human language* (pp. 113–137). Berlin: Walter de Gruyter.
- Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, 66(2013), 30–54.
- Rogers, J. (2003). wMSO theories as grammar formalisms. *Theoretical Computer Science*, 293, 291–320.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926–1928.
- Seki, Y., Suzuki, K., Osawa, A. M., & Okanoya, K. (2013). Songbirds and humans apply different strategies in a sound sequence discrimination task. *Frontiers in Psychology*, 4, 447.
- Yim, O., & Ramdeen, K. T. (2015). Hierarchical cluster analysis: Comparison of three linkage measures and application to psychological data. *The Quantitative Methods for Psychology*, 11, 8–21.