Towards a rationalist theory of language acquisition

Edward Stabler STABLER@UCLA.EDU
Department of Linguistics
University of California, Los Angeles
Los Angeles, CA 90095-1543, USA

Editors: Alexander Clark, Makoto Kanazawa and Ryo Yoshinaka

Abstract

Recent computational, mathematical work on learnability extends to classes of languages that plausibly include the human languages, but there is nevertheless a gulf between this work and linguistic theory. The languages of the two fields seem almost completely disjoint and incommensurable. This paper shows that this has happened, at least in part, because the recent advances in learnability have been misdescribed in two important respects. First, they have been described as resting on 'empiricist' conceptions of language, when actually, in fundamental respects that are made precise here, they are equally compatible with the 'rationalist', 'nativist' traditions in linguistic theory. Second, the recent mathematical proposals have sometimes been presented as if they not only advance but complete the account of human language acquisition, taking the rather dramatic difference between what current mathematical models can achieve and what current linguistic theories tell us as an indication that current linguistic theories are quite generally mistaken. This paper compares the two perspectives and takes some first steps toward a unified theory, aiming to identify some common ground where 'rationalist' linguistic hypotheses could directly address weaknesses in the current mathematical proposals.

Keywords: learnability, syntactic concept, syntactic category

> "What is at stake is... the very foundation on which the current theories are built." (Kracht, 2011, vii)

This paper will compare two theories that differ explicitly in their assumptions about what a human language is, exploring how this difference leads to different conceptions of language learning. At first, it might seem that one view has the 'empiricist' idea that linguistic concepts are sets of perceived data, while the other, more 'rationalist' or 'nativist' theory has intensionally characterized linguistic concepts in compositionally interpreted structures. The different theories can then seem to have no common ground; they are not talking about the same things. Their learning problems differ accordingly. The empiricist has an easy learning theory: the empiricist learner collects data points and observes simple relations among those points. The rationalist learner, on the other hand, must somehow bring data to bear on hypotheses that are fundamentally 'deeper', not being claims about percepts at all. For the rationalist, learning must involve a kind of leap to conclusions that go well beyond the evidence, and so it is sometimes compared to 'triggering'. The recognition of a syntactic concept could be rather like the stickleback's triggered recognition of a mate or rival based on an accidental connections with shape and coloration (Tinbergen, 1952).

It seems that empiricist and nativist conceptions of language and learning could hardly be more different.[1] But when the situation is considered more carefully, it is clear that both sides of the debate have taken aim at straw men. In fact, the empiricist-leaning account is not really empiricist in the naive way just mentioned, and the rationalist-leaning account has good reasons for 'deep' analyses and is not uninterested in how the learner's evidence relates to the acquired concepts. It is still hard to see the outlines of a happy resolution coming into view, but the empiricist is right that the data must be carefully marshalled to get us to the right concepts, and recent proposals surveyed in §1 significantly advance our understanding of how this must work. And the rationalist as described in §2 is right that linguistic concepts must go well beyond the data, that successful identification depends on much more than what perceived evidence entails, and that some of the factors essential for success are accidental. Weaknesses of current mathematical proposals are directly addressed by recent linguistic hypotheses, and here we find the common ground between the two traditions. Some first ideas for linguistic theory are translated into the language of learnability theory in §3, and assessed more carefully in the appendix. These proposals have a 'rationalist' character, since they involve restricting the hypothesis space.

1. String sets and operations on them

Clark (2009, 2014) and related work (Clark et al., 2008; Yoshinaka, 2011; Leiss, 2014) proposes that the syntactic concepts associated with context free grammars (CFGs) are maximal string set/context set pairs, defining some classes of CFLs that can be identified in the limit by a learner with access to positive examples and membership queries, and these ideas are then extending to multiple context free grammars (MCFGs). This section provides a very brief review of some the main ideas of this research, presupposing familiarity with CFGs and MCFGs. The reader should consult cited sources for introductions.

Given any finite, nonempty alphabet Σ, and strings $u, v, w \in \Sigma^*$, define the operation that wraps a pair of strings around another string $\langle u, v\rangle \odot w = uwv$, and extend the operation pointwise to sets of contexts and sets of strings. Then for any language $L \subseteq \Sigma^*$ and any strings $S \subseteq \Sigma^*$ and contexts $C \subseteq \Sigma^* \times \Sigma^*$:

$$\begin{array}{rcl} S^{\triangleright} & = & \{(u,w) | \{(u,w)\} \odot S \subseteq L\} \\ C^{\triangleleft} & = & \{v | C \odot \{v\} \subseteq L\}. \end{array}$$

It is easy to show that these two operations form a Galois connection: $A \subseteq C^{\triangleleft}$ iff $C \subseteq A^{\triangleright}$. The proposal then is that $\langle A, C\rangle$ is a syntactic concept of L iff $A = C^{\triangleleft}, C = A^{\triangleright}$. Letting $\mathcal{C}$ be the set of these syntactic concepts of L with the ordering

$$\langle S_1, C_1\rangle \leq \langle S_2, C_2\rangle \text{ iff } S_1 \subseteq S_2,$$

[1] Clark and Lappin (2011, p163) say: "Representations of language that posit deep hidden parameters render learning intractable... Rather than constituting an argument for strong domain-specific learning priors and biases, they indicate that shallower, observationally accessible representations offer more hopeful systems for language acquisition." Berwick et al. (2013, p33) say: "[Clark & Eyraud] develop an approach that fails even for the simplest examples and completely avoids the original problem... There seems to be no way to remedy the irrelevance of CE's proposal while keeping to anything like their general approach."

$\langle \mathcal{C}, \leq \rangle$ is a complete lattice. By the Myhill-Nerode theorem, $\mathcal{C}$ is finite iff L is regular, and in that case we can write down a grammar corresponding to the lattice as follows. Letting $S_1 S_2 = \{uv | u \in S_1, v \in S_2\}$ as usual, define a product for concepts

$$\langle S_1, C_1 \rangle \circ \langle S_2, C_2 \rangle = \langle (S_1 S_2)^{\triangleright\triangleleft}, (S_1 S_2)^{\triangleright} \rangle.$$

Then, whenever we have $A, B, D \in \mathcal{C}$ with $A \circ B \leq D$, we can have $D \rightarrow AB$. And whenever $A = \langle S, C \rangle$ with $w \in S \cap \Sigma$, we can have $A \rightarrow w$. These rules will generate L.

We could treat non-regular languages with a similar strategy if it would suffice somehow to use only finite $K \subset \Sigma^*$ and finite $F \subset \Sigma^* \times \Sigma^*$. Let's say that a CFG has the f-finite context property (f-FCP) iff for every category A, there is a $C_A \subset \Sigma^* \times \Sigma^*$ such that $|C_A| \leq f$ and $C^{\triangleleft} = \{s \in \Sigma^* | A \Rightarrow^* s\}$. And a CFG has the k-finite kernel property (k-FKP) iff for every category A, there is a $K_A \subset \Sigma^*$ such that $|K_A| \leq k$ and $K_A^{\triangleright} = \{s \in \Sigma^* | A \Rightarrow^* s\}$.[2] For example, a standard grammar of $a^n b^n$ has a kernel of size 6 (Clark, 2009, Fig.2):

$$K = \{a, b, ab, aab, abb, aabb\}.$$

This K is very considerably smaller than Σ^*. For languages with finite kernels, a learner can simply collect the kernel sentences and enough contexts to use them. Given any finite kernel K and contexts F, define a context free grammar with start symbols Z as follows:

$$\begin{array}{rcl}
G^k(K, F, L) = \langle \Sigma, V, \rightarrow, Z \rangle & \text{where} & \\
V & = & \{S \subseteq K | \ |S| \leq k\} \\
S \rightarrow X & \text{iff} & \text{either } X \in \Sigma^{\epsilon} \text{ and } (S^{\triangleright} \cap F) \odot a \subseteq L, \\
& & \quad \text{or } X = BC \text{ and } (S^{\triangleright} \cap F) \odot BC \subseteq L \\
Z & = & \{A | A \subseteq L\}.
\end{array}$$

Note that determining the rules and start categories involves checking whether certain finite sets are subsets of L.

Algorithm 1: Primal learner for k-FKP (Yoshinaka, 2011)

```
Input: sample strings from $L : w_1, \ldots, w_n$; membership oracle for $L$ to calculate $G^k$
$D := \emptyset$;  $K := \emptyset$;  $F := \emptyset$;
for $i = 1$ to $n$ :
    $\hat{G} := G^k(K, F, L)$
    $D := D \cup \{w_i\}$;  $F :=$ all contexts of substrings of $D$
    if $D \not\subseteq L(\hat{G})$ :  $K :=$ all substrings of $D$
return $G^k(K, F, L)$
```

As Clark et al. (2008, p2715) remark, with this representation of the language, if the learner can get answers to membership questions from a 'teacher', an 'oracle', learning becomes trivial. Having collected relevant substrings and contexts, checking on their combinations with the oracle, the learner can "simply write down" a grammar for them. Clark (2010b) and Yoshinaka (2011) propose a 'primal' learner that collects a set K of substrings up to a size bound k, using contexts to determine which rules will be in the language (Algorithm 1).

[2] These definitions follow Yoshinaka (2011, p434), but cf. Clark et al. (2008, p2720), Leiss (2014, §2.3).

This simple learner will identify any k-FKP language in the limit. A 'dual' strategy builds a context set F up to some size bound f, using strings to determine which rules will be in the language. All the learning algorithms in the papers cited in this section are similarly simple. Algorithms of this kind can also be extended to MCFGs (Yoshinaka, 2010). Instead of strings in Σ^* and contexts in $(\Sigma^*)^2$, with a k-MCFG, categories can derive i-tuples in $(\Sigma^*)^i$ for $i \leq k$, with contexts in $(\Sigma^*)^{i+1}$. And when these languages have the analogous finite kernel and finite context properties, similar distributional grammar inference methods can succeed.

Previewing later conclusions, we can already see that these recent learnability results are linguistically important. These models do not pretend to provide a reasonable model of human-like acquisition, but they are relevant to understanding how that could work. *First,* the required fit between the evidence and the acquired linguistic concepts must be something like what we see in these models. This is established by the foundational work on the interpretation of the grammars in the hypothesis space.[3] *Second,* when we reflect on the very general issues mentioned at the outset, we see that the proposals here are not naively empiricist. While the finite sets of strings K, F might be regarded as sets of percepts, when the learner succeeds it is because the target language has the very special structural property, k-FCP. Because of that property, these very small sets can act as 'triggers' for the grammar of the infinite language. Rationalists may overplay the analogy between fish and human language learners, but Tinbergen is right about both when he says that they are responding to "a few characteristics of an object rather than to the object as a whole."[4] *Finally,* it is illuminating to note that from any standard linguistic standpoint, k-FCP or any similar property would certainly be regarded as *accidental.* That is, k-FCP is a property that a language may have or not regardless of what the structures of sentences are. Possession of the k-FCP depends, for example, on the extent to which different categories and their contexts are homophonous, while no syntactic operation depends on any such thing.

Turning now to problems we would like to correct, three main issues limit the linguistic interest of this work. *First, the hypothesis space is too large.* These models fail to predict most basic facts about human language. If we ask which properties of human language would be entailed if human learners used Algorithm 1 or something like it, it is not easy to find any that would merit mention in any introduction to linguistics. The hypothesis classes of CFLs or MCFLs with the k-FKP are (possibly too small but also definitely) much too large. We return to this in §3 below. *Second, the learner's weakly adequate grammars are strongly inadequate.* That is, the proposed learners will converge on grammars that generate the target language, but the learners' grammars are not compact, missing many generalizations. The learners' grammars are so unlike traditional grammars that it is difficult to see how they would allow human language to play its role in thought and reasoning. The correspondence between structure, semantic value, and inferential role may

[3] The positive learnability results establish the value of the methodological strategy of first studying the grammars and their connection to the data. This is not a new idea (Chomsky, 1975, p15), but here we see how the foundational work determines what the learning problem is. As in computing quite generally, running programs on particular examples to see what they do is usually much less valuable than carefully considering what needs to be achieved and what kinds of computations could achieve that.

[4] Tinbergen (1952) also notes that while mammals can adapt appropriately to more situations, given "the affinity of mammals to lower vertebrates, one expects to find an innate base beneath the plastic behavior of mammals." That is what we are looking for.

not be as tight in human languages as it is in propositional logic, but a model in which five word sentences are parsed in hundreds of ways does not look promising.[5] *Third, these models are infeasible.* Once the learning problem is well-understood, we can hope that the membership oracle can be replaced with indirect statistical evidence, but even with an oracle, the hypothesized grammars can be hundreds of times larger than the sample, and compute times are large. In senses that the literature has precisely defined, these models are polynomial in the size of the input sample, but for large samples that is not good enough for feasibility.[6] Worries about efficiency should mainly be ignored at early stages of studies like this, but they are worth noting here because of their obvious connection to the two previously mentioned problems. The complexity comes in part from generating structures that, plausibly, no human would ever consider. The idea of reducing the size of hypothesized grammars after they are built (Clark, 2010a, ex.1) does not seem like a good one: redundancies exact a cost in the inference steps that build and use them, in every loop of the inference method, and finding redundancies in these grammars is generally undecidable (Chomsky and Schützenberger, 1963). Approximate methods can be used (Brabrand et al., 2007; Schmitz, 2007), but it is hard to see how they would be part of an insightful approach to language acquisition. I think the problem is in the representation of the languages.

2. Categorized expressions and operations on them

It is difficult to compare the distributional perspectives on language with more traditional ones, partly because the 'more traditional' perspectives are so diverse. Keenan and Stabler (2003) present one way of getting at some of the most basic 'traditional' ideas about structure. They begin by assuming that the rules of a grammar G operate not on strings but on categorized strings. Given alphabet Σ and categories Cat, the possible expressions are $\Sigma^* \times Cat$. The lexicon Lex is a finite set of these elements, and the rules F of the grammar are partial functions from (tuples of) possible expressions to possible expressions. The language $L(G)$ is the closure of Lex with respect to F. The string yield of any category $c \in Cat$, $Str(c) = \{s | \langle s, c\rangle \in L(G)\}$.

This framework is so flexible that a very wide range of grammar formalisms can be translated into it. For example, consider an MCFG in which all rules have the following form, for some $n \geq 0$:

$$A_0(\alpha_1, \ldots, \alpha_{r(A_0)}) \leftarrow A_1(x_{1,1}, \ldots, x_{1,r(A_1)}) \ldots A_n(x_{n,1}, \ldots, x_{n,r(A_n)})$$

where each $\alpha_i \in (\Sigma \cup X)^*$, each $x_{i,j}$ is a variable, and each variable occurs exactly once on the right and at most once on the left.

Let's call the rules with $n = 0$ the lexical rules, and assume that (i) the non-lexical rules have no terminal symbols on their left sides and (ii) no two non-lexical rules have right sides that are equivalent up to renaming of variables. These conditions do not affect the

[5] In categorial and type-logical grammars too there can be many derivations of each string, but in those grammars, there is still something special about the division between subject and predicate in a sentence like [three mathematicians in ten][derive a lemma] (Steedman, 2000). Special syntactic, prosodic and interpretive effects of that boundary can be predicted. For the distributional learners, on the other hand, the spurious derivations have no connection to anything.

[6] Showing their learner is polynomial in the size of the sample, Clark et al. (2008, p2722) emphasize "this is not a strong enough result."

expressive power of MCFGs, as standardly interpreted, and these grammars have an easy translation into the kind of system studied by Keenan and Stabler (2003). Let rules with $n = 0$ specify $Lex = \{\langle(\alpha_1, \ldots, \alpha_{r(A)}), A\rangle | A(\alpha_1, \ldots, \alpha_{r(A)}) \leftarrow\}$. The rules with $n > 0$ can be regarded as specifying a structure building function that maps a string substitution instance of any right side to the corresponding instance the left side:

$$\langle(x_{1,1}, \ldots, x_{1,r(A_1)}), A_1\rangle, \ldots, \langle(x_{n,1}, \ldots, x_{n,r(A_n)}), A_n\rangle \mapsto \langle(\alpha_1, \ldots, \alpha_{r(A_0)}), A_0\rangle.$$

With this interpretation, the strings of the start category $Str(S)$ are exactly the language of the MCFG under its usual interpretation (Seki et al., 1991).

Lifting any function on $L(G)$ to apply coordinatewise to tuples in $L(G)^*$, and then pointwise to sets of expressions or tuples of expressions, an automorphism is a bijection on the language that leaves the functions F unchanged. That is, bijection $h : L(G) \rightarrow L(G)$ is an automorphism of $(L(G), F)$ iff for every $r \in F$, $h(r \upharpoonright L(G)) = r \upharpoonright L(G)$. It is easy to prove that the identity map on $L(G)$ is an automorphism of every G. And so is h^{-1} whenever h is, and so is the composition $g \circ h$ whenever g and h are. So the set of automorphisms is a group. We extend any automorphism h pointwise to map subsets of $L(G)$ to the sets of their values. And we extend any such h to $L(G)^*$, so that $h(a_1, \ldots, a_n) = \langle g(a_1), \ldots, g(a_n)\rangle$. Then the invariants of G are the fixed points of the automorphisms of G, so extended. And let's say that expressions s, t are structurally equivalent, iff $h(s) = t$ for some automorphism h. It is easy to see that this relation partitions $L(G)$.

With these definitions, it is possible for an expression of category c to be structurally equivalent to an expression of another category. And although the categories partition the language, that partition can fail to be a congruence. That is, it can happen that for s, t of the same category, s is in the domain of rule r but t is either not in the domain of the rule or is in the domain but the value of the rule applied to s does not have the same category as the rule applied to t. Keenan and Stabler (2003) discuss cases where linguists have proposed grammars like this. But the grammars linguists need have a number of more basic universal properties, including things like the following: Grammars for human languages are category-functional in the sense that the category of a complex is a function of the categories of its immediate constituents (p153); Expressions which are syntactically invariant are also semantically invariant (p165); If some automorphism permutes two distinct lexical items s, t, then any rule applied to s yields a different value from the rule applied to t (p164); and if s is a proper constituent of t, then $s \neq t$ (p160).

None of these universals say anything about the pronounced string parts of any expression. When we say two expressions have the same structure, or that a rule applies to an expression or not, the associated string values, considered by themselves, are irrelevant. de Saussure (1907) is famous for noting the arbitrariness of signs; Pullum and Zwicky (1988) argue that nothing in syntax needs to see phonetic properties; Newmeyer (2005, pp4-6) includes in his list of universals the claim that no syntactic process can be sensitive to the segmental phonology of the lexical items undergoing the process, and so no language has segmental/phonological conditions on word order. Stabler and Keenan (2007) propose ways to identify basic patterns of recursion, predication and modification in grammars where the categories are not labeled, possibly giving us access to clauses, predicates, arguments, and modifiers, but it is not immediately clear how to get this from string sets.

Consider, for example, the English verb *have* from traditional and distributional perspectives. This is a very special verb in English, so it is plausible that in any reasonable English grammar, the verb *have* is invariant: no other word could take the place of this verb in all derivations, so no automorphism that leaves the structure building rules unchanged can map any other word to this one. On the other hand, if we ask what can go into the pronounced context (*I,eaten*), we see that *have* occurs there along with infinitely many other things like *was, got, know she has,*.... In most American dialects, the verb *have* is homophonous with *halve*, and so if we were paying attention to pronunciation rather than spelling, both of those different verbs would appear as the same element in contexts like (*I,the orange*), and infinitely many other expressions could appear there too. What a distributional grammar will not do is tell us whether any word in (*I,the orange*)$^{\triangleleft}$ is the 'same word', in the traditional sense, as any word in (*I,eaten*)$^{\triangleleft}$. Without such an identification, it is difficult to see how the analysis could underpin reasoning about patterns of predication, let alone semantic and pragmatic assessment.

3. Towards a rationalist theory of the learner's data

Most of the proposed linguistic universals in the literature are much less abstract than the ones proposed by Keenan and Stabler (2003). They are also more controversial (Evans and Levinson, 2009), and similarly difficult to connect to distributional clues. To mention just a few ideas that could be relevant to a learner:

(H1) *Fixed universal categories, clause structure, cartography.* Many linguists have assumed that the set of categories is fixed and finite – see for example Chomsky (1981, p11), Pinker (1982, p672). Similar claims can be found in recent literature in the 'cartographic' tradition, for example, where it is sometimes claimed that not only are the categories fixed and finite, but their 'underlying' order in the clause is also fixed, with deviations caused by movements. These proposals are preliminary, of course, but one critique is that, even with a fixed underlying order of all elements, the movement options allow all possible orders to be derived, so these claims are difficult to falsify. See for example van Craenenbroek (2009) and references cited there.

(H2) *Limits to selection.* Pesetsky (1995, §6.1.5.4) argues that a predicate can select a subject and no more than two obligatory internal arguments. Less specific claims have been around in the literature for some time. Chomsky (1981, p11) says "subcategorization frames and the like are narrowly limited in variety," but never spells out what these limitations are. Compare similar claims in Pinker (1982, p672).

Abney (1987, pp64ff) argues that 'functional' heads (complementizer, tense, preposition, determiner,...) typically select a fixed category (tense phrases, verb phrases, determiner phrases, noun phrases,..., respectively); they usually cannot be separated from their complements; they are often phonologically dependent; and they denote in higher types than nouns or verbs, typically "regulating or contributing to the interpretation of their complement."

(H3) *Bounded clausal embedding.* Wexler and Culicover (1980) propose that every natural language grammar can be identified by 'degree 2' structures, structures with no more than two embedded clauses, since all variations in phrase structure and in patterns of movement can be seen in structures of that size. Lightfoot (1989) argues that even less may suffice; a rich enough assessment of degree 1 data may suffice.

(H4) *Parameters.* Many linguists think there is a finite number of fixed, universal 'parameters' of language variation, so that languages fall into one of a finite number of basic types. For example, Roberts (2012) suggests that in some languages, heads are always (underlyingly) final in their phrase; in some languages heads are all initial; in some languages only the verb phrases are head-final, and so on. Some linguists think that this syntactic variation can be attributed to the properties of functional elements in the extended projections of these heads. Perhaps even arbitrarily many fine-grained micro-parameters vary from each speaker to the next (Kayne, 2000).

All of these ideas refer to grammatical categories and structures which the learner will not, initially, be able to identify in sequences of words. Chomsky observes this general problem for 'universal grammar' (UG):

> . . . in the case of UG . . . we want the primitives to be concepts that can plausibly be assumed to provide a preliminary, prelinguistic analysis of a reasonable selection of presented data, that is, to provide the data that are mapped by the language faculty to a grammar. . . It would, for example, be reasonable to suppose that such concepts as "precedes" or "is voiced" enter into the primitive basis. . . But it would be unreasonable to incorporate, for example, such notions as "subject of a sentence" or other grammatical notions, since it is unreasonable to suppose that these notions can be directly applied to linguistically unanalyzed data. (Chomsky, 1981, p10)

Assuming that the set of possible human grammars can be bounded to a finite number n, he suggests: "it is quite possible that there exists a finite set of sentences S such that systematic investigation of S will suffice to distinguish the n possible grammars" (Chomsky, 1981, p11). Unfortunately no details are provided about how this could work.

One common idea is that the data needs to be enriched somehow with semantic values, or clues to semantic values. Perhaps the syntax could be 'bootstrapped' from the semantics. While rich semantic data could trivialize the learning problem, as Clark and Lappin (2011, p68) point out, it is difficult to see how a typical learner could get useful semantic clues at the early stages of language acquisition (Gleitman et al., 2005). It is perhaps more natural to assume that the syntax is mastered first, and leads the child to be able to conceive things that would otherwise be impossible – a kind of 'syntactic bootstrapping'. I think it is safe to say that preliminary explorations of these ideas have not yet led to any big breakthrough in learnability.

The distributional learners suggest some new possibilities. I think there are some combinations of language universals like the ones mentioned above that could restrict the hypothesis space in principled ways, simplifying the distributional learning problem. The kind of reasoning about substitution classes found in Harris (1946) is also used in contemporary syntax, but without the assumption that it exhausts the relevant considerations. The introductory text Sportiche et al. (2013, pp52ff), for example, uses distributional considerations to probe the structure of sentences like these:

she will put it there
she will put this one there then
she will put it on this one then
this one will put it there then
this girl in the red coat will put a picture of Bill on your desk before tomorrow
she will put a picture of Bill on your desk before tomorrow
this girl in the red coat will put it on your desk before tomorrow
this girl in the red coat will put a picture of Bill there before tomorrow
this one will put a picture of Bill on your desk before tomorrow
this girl in the red coat will put a picture of Bill on it before tomorrow
this girl in the red one will put a picture of Bill on your desk before tomorrow

Typical linguistic students can act as their own oracles to explore more variations of these sentences, and they can use semantic intuitions. But when we look at how these examples are handled by Algorithm 1 (or any of the other alternatives mentioned in §1), the basic patterns in this data are not found easily. One reason is that unnecessarily long contexts are considered. This particular collection of data is actually designed to get the student to see that the obligatory elements in this kind of sentence are just 5 in number. As (H2) suggests, the first sentence is a case where the verb selects a subject and two internal arguments. Since the verb, auxiliary, subject and two internal arguments can each be expressed with a single word, 5 words, 5 heads is enough. In all the other cases, we have longer sequences playing the roles of single words in the shorter sentences. Consequently, the learner should never, in English or any other language, need to consider a word context of length 16 like

$$C\text{=(this girl in the red coat will put a, of Bill on your desk before tomorrow)}$$

This is because $C^{\lhd} = (\textit{she will put a,there})^{\lhd}$.

Recent versions of (H1) often list the order of very elaborate clausal structures, but considerations like (H2) remind us that many of those elements are optional. Simplifying and combining these ideas, there is plausibly a maximum number of obligatory phrases in any clause: $e_1, e_2, \ldots, e_n$, with a maximum number of words $\ell(e_i)$ required to form each element e_i. In that case, the clausal context of any substring w that occurs in $e_i = uwy$ can be given by a context of length

$$\ell(e_1) + \ldots + \ell(e_{i-1}) + \ell(u) + \ell(y) + \ell(e_{i+1}) + \ldots + \ell(e_n).$$

and if no more than degree 2 or even degree 1 data is all that is ever needed, then we could get a principled bound on the needed context sizes. Algorithm 1 imposes a bound k on the size of context sets, but no bound on the lengths of particular contexts. It is plausible that, for certain kinds of languages, both can be bounded. This idea and some other similar bounds on grammar complexity are explored in the appendix http://www.linguistics.ucla.edu/people/stabler/icgiApp.pdf.

Acknowledgments

I would like to acknowledge support for this project from the National Science Foundation (NSF grant IIS-1125423).

References

Steven P. Abney. *The English Noun Phrase in its Sentential Aspect.* PhD thesis, Massachusetts Institute of Technology, 1987.

Robert C. Berwick, Noam Chomsky, and Massimo Piattelli-Palmarini. Poverty of the stimulus stands: Why recent challenges fail. In Massimo Piattelli-Palmarini and Robert C. Berwick, editors, *Rich Languages from Poor Inputs.* Oxford University Press, NY, 2013.

Claus Brabrand, Robert Giegerich, and Anders Møller. Analyzing ambiguity of context-free grammars. In *Proceedings of the 12th International Conference on Implementation and Application of Automata, CIAA'07*, 2007.

Noam Chomsky. *Reflections on Language.* Pantheon, NY, 1975.

Noam Chomsky. *Lectures on Government and Binding.* Foris, Dordrecht, 1981.

Noam Chomsky and Marcel-Paul Schützenberger. The algebraic theory of context-free languages. In P. Braffort and D. Herschberg, editors, *Computer Programming and Formal Systems*, pages 118–161. North-Holland, Amsterdam, 1963.

Alexander Clark. A learnable representation for syntax using residuated lattices. In *Proceedings of the 14th Conference on Formal Grammar*, 2009.

Alexander Clark. Disibutional learning of some context-free languages with a minimally adequate teacher. In *Proceedings of the International Conference on Grammatical Inference, ICGI'10*, 2010a.

Alexander Clark. Learning context free grammars with the syntactic concept lattice. In José Sempere and Pedro Garcia, editors, *Grammatical Inference: Theoretical Results and Applications. Proceedings of the International Colloquium on Grammatical Inference*, pages 38–51. Springer, 2010b.

Alexander Clark. The syntactic concept lattice: Another algebraic theory of the context-free languages? *Journal of Logic and Computation*, Forthcoming, 2014.

Alexander Clark and Shalom Lappin. *Linguistic Nativism and the Poverty of the Stimulus.* Wiley-Blackwell, NY, 2011.

Alexander Clark, Rémi Eyraud, and Amaury Habrard. A polynomial algorithm for the inference of context free languages. In *Proceedings of the International Conference on Grammatical Inference, ICGI 2008*, 2008. Video http://videolectures.net/icgi08_clark_pai/.

Ferdinand de Saussure. *Premier Cours de Linguistique Générale.* Pergamon, NY, 1907. 1996 French-English edition, with English translation by George Wolf, edited by Eisuke Komatsu.

Nicholas Evans and Stephen Levinson. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–448, 2009.

Lila R. Gleitman, Kimberly Cassidy, Rebecca Nappa, Anna Papafragou, and John C. Trueswell. Hard words. *Language Learning and Development*, 1(1):23–64, 2005.

Zellig S. Harris. From morpheme to utterance. *Language*, 22:161–183, 1946.

Richard S. Kayne. Microparametric syntax: Some introductory remarks. In *Parameters and Universals*, pages 3–9. Oxford University Press, Oxford, 2000.

Edward L. Keenan and Edward P. Stabler. *Bare Grammar*. CSLI Publications, Stanford, California, 2003.

Marcus Kracht. *Interpreted Languages and Compositionality*. Studies in Linguistics and Philosophy 89. Springer, Berlin, 2011.

Hans Leiss. Learning CFGs with the finite context property: A correction of A. Clark's algorithm. In *Proceedings, Formal Grammar*, Tübingen, 2014.

David Lightfoot. The child's trigger experience: degree-0 learnability. *Behavioral and Brain Sciences*, 12:321–375, 1989.

Frederick J. Newmeyer. *Possible and Probable Languages: A Generative Perspective on Linguistic Typology*. Oxford University Press, Oxford, 2005.

David Pesetsky. *Zero Syntax: Experiencers and Cascades*. MIT Press, Cambridge, Massachusetts, 1995.

Steven Pinker. A theory of the acquisition of lexical interpretive grammars. In Joan Bresnan, editor, *The mental representation of grammatical relations*. MIT Press, Cambridge, Massachusetts, 1982.

Geoffrey K. Pullum and Arnold M. Zwicky. The syntax-phonology interface. In Frederick J. Newmeyer, editor, *Linguistics: The Cambridge Survey*, pages 255–280. Cambridge University Press, NY, 1988.

Ian Roberts. Macroparameters and minimalism. In C. Galves, S. Cyrino, R. Lopes, F. Sandalo, and J. Avelar, editors, *Parameter Theory and Linguistic Change*, pages 320–335. Oxford University Press, NY, 2012.

Sylvain Schmitz. Conservative ambiguity detection in context-free grammars. In *Proceedings of the 34th International Conference on Automata, Languages and Programming, ICALP'07*, LNCS 4596, pages 692–703. Springer-Verlag, Berlin, Heidelberg, 2007. ISBN 3-540-73419-8, 978-3-540-73419-2.

Hiroyuki Seki, Takashi Matsumura, Mamoru Fujii, and Tadao Kasami. On multiple context-free grammars. *Theoretical Computer Science*, 88:191–229, 1991.

Dominique Sportiche, Hilda Koopman, and Edward Stabler. *An Introduction to Syntactic Analysis and Theory*. Blackwell, Oxford, 2013.

Edward P. Stabler and Edward L. Keenan. Universals across languages. Workshop on Model Theoretic Syntax, ESSLLI'07, 2007.

Mark Steedman. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31(4):649–689, 2000.

Nikolaas Tinbergen. The curious behavior of the stickleback. *Scientific American*, pages 22–26, 1952. December.

Jeroen van Craenenbroek. Introduction. In Jeroen van Craenenbroek, editor, *Alternatives to Cartography*, pages 1–14. Mouton de Gruyter, NY, 2009.

Kenneth Wexler and Peter W. Culicover. *Formal Principles of Language Acquisition.* MIT Press, Cambridge, Massachusetts, 1980.

Ryo Yoshinaka. Polynomial-time identification of multiple context-free languages from positive data and membership queries. In *Proceedings of the 10th International Colloquium on Grammatical Inference*, LNCS 6339, Berlin, 2010. Springer-Verlag.

Ryo Yoshinaka. Towards dual approaches for learning context-free grammars based on syntactic concept lattices. In *Proceedings of the 15th International Conference on Developments in Language Theory*, LNCS 6795, pages 429–440. Springer-Verlag, 2011.