Grammars volume 7(2004), 57–68

Computational Grammar Induction for Linguists

Pieter Adriaans
ILLC, University of Amsterdam, The Netherlands
Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands
E-mail: pietera@illc.uva.nl

Menno van Zaanen
ILK, Tilburg University, The Netherlands
5000 LE, Tilburg, The Netherlands
E-mail: mvzaanen@uvt.nl

1 Introduction

In general a grammar describes a (possibly infinite) set of sentences with a finite structural description. Computational Grammar Induction (CGI) deals with the creation of computational models for identification of these infinite sets on the basis of a finite set of examples. CGI is a field in its own right, with its own internal research questions, many of which have no direct impact on the study of human language. Yet it is clear that computational models created by the CGI community might be of interest to the linguistic community because human language after all appears to be an infinite set, the description of which is learned efficiently in a relative short time. There are various domains in which learnability of a language might be interesting for linguists: e.g., first language acquisition, second language acquisition, or the automatic extraction of grammars from corpora.

In this article we will focus on first language acquisition with some suggestions for extensions to other areas.[1] with some suggestions for extensions to other areas. Up till now there has been surprisingly little cross-fertilization between the field of linguistics and that of Computational Grammar Induction. There are various reasons for this situation, for example, there is a strong interest in the descriptive study of language in linguistics on the one hand and a strong focus on abstract language models in the CGI community on the other hand.

[1] There is some discussion about the complexity of human languages, but the general consensus is that they are at least context-free and with high probability context-sensitive [Huybrechts, 1984; Shieber, 1985]. In this article we will therefore mainly focus on learning context-free languages as a sort of first approximation of human languages. There has been little research into the learnability of context-sensitive languages. It is also unlikely that the full power of the context-sensitive class of languages is necessary for the description of natural language. Context sensitive constructions seem to exist but but in rather isolated cases and with theoretically infinite, but in practice rather shallow forms of recursion.

A factor that has certainly contributed this lack of cross-fertilization is the famous conjecture of Chomsky that the efficiency of human language acquisition can only be explained on the basis of an innate Universal Grammar (UG). The UG discussion, which is in fact a revival of the ancient philosophical debate of rationalism (Descartes) versus empiricism (Hume), has attracted a lot of attention with surprisingly little tangible results.

At this stage of writing it has become clear that the study of human language acquisition requires a huge interdisciplinary effort. Co-operation between linguists, psychologists, audiologists, neuro-physiologists, cognitive scientists, and computer scientists will be necessary to create better models of language acquisition. There will be also a huge empirical effort involved. For example, the creation of elementary annotated video sequences to study the language development of children is an investment that easily amounts to several hundreds of person-years. In order to set up an experiment that allows conclusions with any statistical significance one needs of a group of 20 children (a target group of 10, a control group of 10), 1 hour of video per child per week, duration of three years, plus annotation: behavioral and dialogical interaction between mother and child, phonetic and orthographic transcription, further linguistic analysis. This is a total effort of ca. 150 person-years. Multiply this by the number of languages and number of target groups (hearing impaired children, bi-lingual children, etc.) one would like to analyze. Add the costs of measuring brain activity, using ERP and fMRI scans. Then observe that some experts think that 1 hour of video per week is insufficient to capture the sudden spurts of increase in linguistic capability that children sometimes display in a couple of days. These examples illustrate that the analysis of human language acquisition is a huge effort that requires international and interdisciplinary co-operation.

It is clear that the understanding of language acquisition is not a matter that can simply be decided in terms of rough philosophical schemes like tabula rasa versus innate ideas. For the moment the only sensible thing to do is postpone the debate until more empirical evidence is available. It is perfectly possible to start to chart a river without knowing where its source lies. Instead of trying to come up with one grand scheme, one should analyze local transformations. Current research focuses on a good understanding of the various stages of language acquisition and tries to understand the complexity of transformations between different stages. In the rest of this article we will first give a birds eye view of Grammar Induction research, then give a sketch of the problems that are encountered when analyzing children's language acquisition and finally discuss the application of CGI techniques to these problems.

2 A birds eye view of CGI

Grammar Induction can be seen as the identification of an infinite structure with a finite structural description on the basis of a finite number of examples. As was already observed by Hume [1909] in the 17th century, such a form of induction cannot be made deductively valid. It is in general not possible to derive a universal rule on the basis of a finite set of examples. The codification of this principle in terms of

Grammar Induction is found in the work of Gold [1967].

Gold presents a game theoretical definition of learnability called *identification in the limit*. There are two parties: a Challenger and a Learner. At the start of the process both parties agree on a class of languages. The Challenger selects a language from this class and presents an enumeration of structures to the Learner. If the Challenger only presents sentences of the language then we speak about learning from *positive information*, if the Learner also gets negative information we are learning from *complete information*. The Learner produces an infinite number of guesses. If *in the limit* the guesses of the Learner stabilize at the right guess then there is a winning strategy for the learner and the language is learnable. It is not surprising that from a linguistic point of view in this deductive paradigm of learnability, no interesting classes of languages are learnable from positive information only.[2]

The crucial theorem is:

Theorem 1 (Gold) *A class G of grammars is not learnable from positive information if G contains all finite languages and at least one infinite language.*

Such a class of languages has what is called *infinite elasticity*: there is an infinite sequence of real inclusions of languages. At any point of the learning process the Learner has an infinite number of options left. One can prove that *finite elasticity* is a sufficient characteristic of classes of languages that are identifiable in the limit. If one combines the fact that all reasonable formal models of the class of natural languages have infinite elasticity with the general view that children do not get negative information from their parents, (however, see [Sokolov and Snow, 1994] for a discussion.) then this seems to give a weak argument in favor of the existence of a Universal Grammar. Without help of innate structures children would not be able to cope with the complexity of the task of learning their language at all. However, as we have argued above, it is too early to decide on the scientific merit of this view.

One could argue that identification in the limit is not really an inductive approach. The Learner simply waits until there is enough information to deduce the nature of the grammar from the data. Also, the mode of presentation of the examples by the Challenger is completely free except for the constraint that any structure has to occur at least once in the limit. In the case of human communication, the dialogue partners are probably more co-operative. Under representative probability distributions, the class of learnable languages is considerably larger. An important task in the context of the application of formal results from the CGI community to natural languages is to come up with a set of probability distributions that properly reflect the bias of sentence production in various communicative systems: dialogues between mother and child, rational dialogues between researchers, dialogues between teachers and pupils.

[2]Gold's results are represented in the following table:

	Complete Information	*Positive Information*
Recursively Enumerable Sets	Not learnable	Not learnable
Decidable Rewriting Systems	Learnable	Not learnable
Super Finite Set	Learnable	Not learnable
Finite Sets	Learnable	Learnable
Finite Class of Finite Sets	Learnable	Learnable

A larger part of the empirical content of a theory of the learnability of natural languages will consist of a better understanding of the universal bias that governs the probability distributions of human communication. Such a theory of universal bias could serve as a counterpart of the notion of a universal grammar.

An initial estimate of relevant probability distributions and alternative learning models can be formulated with the help of algorithmic complexity theory. One of the bigger intellectual achievements of computer science in the 20th century is the formulation of an adequate theory of induction based on a deep analysis of the relation between the notion of the computational complexity of an object and its a priori probability. The basic insight is that objects that are easier to compute have a higher probability. This idea is relevant for language learning because it helps us to formulate a first approximation of co-operative linguistic behavior: if one wants to teach a language, it appears reasonable to select examples that are easy to analyze first. Using algorithmic complexity theory one can give adequate answers to questions like: Given a text T and two grammars G_1 and G_2 are we able to approximate which grammar has the highest probability of generating T, i.e., $\text{argmax}(P(G_1|T), P(G_2|T))$?

Building on related ideas of Solomonoff, Horning [1969] proved that so-called probabilistic context-free grammars[3] can be learned from positive data only. This result removes the sting of the strict unlearnability results of Gold.

Based on these early findings one can identify three research lines that extend more or less to the present with regrettably little mutual exchange of ideas.

The first line is the extension of the recursion theoretic approach of Gold. Important results in this area are Angluin's work on tell-tale sets [Angluin, 1980],[4] her work on identification with the help of more powerful clues [Angluin et al., 1997] like membership queries[5] and equivalence queries.[6] Even equivalence queries do not guarantee exact identification of context-free grammars in polynomial time. An important negative result is given by Pitt and Warmuth [1988] who prove that (modulo a polynomial time transformation) context-free grammars are as hard to predict as certain cryptographic predicates are to compute.

Given these quite disappointing negative results, the recursion theoretic line of research has focused on learning subclasses of context-free grammars. We mention a few: k-bounded context-free grammars [Angluin, 1987], (k)-reversible languages [Angluin, 1982; Sakakibara, 1992] and simple deterministic languages [Yokomori, 2003; Ishiszaka, 1990]. None of these subclasses can claim to have any specific linguistic importance. This is less true of the interesting results of Kanazawa [1995] concerning learnability of certain classes of categorial grammars.

The second line of attack is taken by a group of researchers that simply try to

[3]The class of probabilistic grammars specifies for each structure a probability which indicates in how far it belongs to the language. An idea that seems to be quite close to the notion of grammaticality in natural languages: almost any structure can be made 'grammatical' in the right context. See the discussion of bootstrapping phenomena below.

[4]Finite subsets of a language that prevent overgeneralization.

[5]Queries that allow the Learner to check whether a certain sentence is in the target language.

[6]Queries that allow the Learner to check whether the current hypothesis concerning the target grammar is correct and return a sentence in the symmetric difference of the hypothesis and the target grammar if this is not the case.

develop efficient learning algorithms that work more or less successfully on real life corpora. An early exponent of this approach is [Wolff, 1977, 2003] with the SP framework, based on information compression and ideas of Solomonoff [1997] and Rissanen. The work of Adriaans [1992] who developed the EMILE framework at the end of the eighties, has culminated in the EMILE 4.1 implementation by Adriaans and Vervoort [2002]. A methodology related to the SP framework, is presented by van Zaanen [2002]: Alignment-Based Learning. Recently van Zaanen and Adriaans [2001] have developed a methodology to compare the efficiency of implementations of Grammar Induction systems using annotated corpora like the Penn-tree bank and the ATIS corpus. Also in this category we find the extensive research effort to come up with efficient algorithms for regular languages that takes place in the ICGI community.[7] This body of work has lead to algorithms that can learn rather complex Deterministic Finite Automata (DFA) with several hundreds of internal states. A good example of state of the art is the so-called blue fringe algorithm [Lang et al., 1998].

A third, and theoretically promising, line of attack of the Grammar Induction problem is the probabilistic approach [Charniak, 1993]. Probabilistic context-free grammars are the only linguistically relevant class of languages that are known to be identifiable in the limit. The algorithm presented by Horning is too complicated to be of any empirical relevance but it suggests interesting further research. Part of these developments took place in the wider context of computational learning theory. Valiant [1984] introduced a completely new learning concept that since then has been named "probably approximately correct" (PAC) learning.

Definition 1 *Let F be a concept class, δ $(0 \leq \delta \leq 1)$ a confidence parameter, ϵ $(0 \leq \epsilon \leq 1)$ an error parameter. A concept class F is* PAC learnable *if for all targets $f \in F$ and all probability distributions P on the sample space U^* the learning algorithm A outputs a concept $g \in F$ such that with probability $(1 - \delta)$ it holds that we have a chance on an error with $P(f \Delta g) \leq \epsilon$, where $f \Delta g = (f - g) \cup (g - f)$.*

The basic idea of PAC learning is that we can minimize the chance of learning something that is wrong without being completely sure that we are right. Valiant applied his theory to Boolean concept learning. Subsequent investigations showed that this technique actually has a much broader field of application [Li and Vitányi, 1991].

Unfortunately PAC learning in its pure distribution free form does not help Grammar Induction much. No linguistically interesting classes of languages are known to be distribution free PAC learnable. This is in itself not very alarming since learning under any distribution is certainly too strong a demand for any form of linguistically relevant form of Grammar Induction. Yet there is more bad news. From empirical observations we know that the distributions that govern human communication are invariably heavy tailed: there is at any stage of the sampling process a considerable probability mass in the set of unseen examples.[8] This is particularly harmful for PAC convergence.

[7] The International Grammar Induction Community, which organizes a two year conference on the subject as well as several Grammar Induction competitions: Abbadingo, Gowachin, and recently, Omphalos.

[8] The distributions of word frequencies in almost all texts seems to follow the so-called Zipf power law: the size of the word frequency class with frequency n is $cn^{-1.5}$, where c is a constant depending

A better understanding of the distributions that govern human communication is necessary. Promising insights can be found in algorithmic complexity theory that allows us to formulate the notion of a priori probability of strings as well as a class of so-called *simple distributions* that might serve as a first approximation of distributions that govern human communication. This has lead to the notion of PAC-s learning: PAC learning under simple distributions [Li and Vitányi, 1991]. A central result is the so-called coding theorem from Levin:

Theorem 2 (Levin)

$$-\log \boldsymbol{m}(x) = -\log P_U(x) + O(1) = K(x) + O(1)$$

Here $\mathbf{m}(x)$ is a universal semi measure,[9] $P_U(x)$ is the universal a priori probability of a binary string x,[10] and $K(x)$ is the prefix Kolmogorov complexity of the string x, i.e., the length of the shortest prefix-free program that generates x on a universal Turing machine.[11]

The importance for a theory of learning is apparent from the following completeness result of Li and Vitányi [1991]:

Theorem 3 *A concept class C is learnable under $\boldsymbol{m}(x)$ iff C is also learnable under any arbitrary simple distribution*[12] *$P(x)$ provided the samples are taken according to $\boldsymbol{m}(x)$.*

Motivated by these ideas Adriaans [2001] formulated the conjecture that natural languages are shallow,[13] i.e., they can be learned from a relatively small set of examples[14] the length of which is logarithmic in the Kolmogorov complexity of the grammar:

Definition 2 *A language G is called shallow if it has a tell-tale set $C \subseteq G$ for which* $\forall s \in C(|s| \leq c \log K(G))$

on the size of the sample. In recent years our understanding of these dimensionless distributions has increased considerably [Faloutsos et al., 1999], but the exact reason why they occur with such regularity in linguistic structures is ill-understood, as is the meaning of the -1.5 exponent.

[9] A recursively enumerable semi-measure μ is called universal if it recursively dominates every other enumerable semi-measure μ', i.e., $\mu(x) \geq c\mu'(x)$ for a fixed constant c independent of x. Levin proved that there is a universal enumerable semi-measure. We fix a universal semi-measure $\mathbf{m}(x)$. The semi-measure $\mathbf{m}(x)$ converges to 0 slower than any positive enumerable function which converges to 0. Of course, $\mathbf{m}(x)$ itself is not recursive.

[10] Defined as : $P_U(x) = \sum_{U(p)=x} 2^{-|p|}$.

[11] The descriptive complexity of a string x relative to a Turing machine T and a binary string y is defined as the shortest program that gives output x on input y: $K_T(x|y) = \min\{|p| : p \in \{0,1\}^*, T(p,y) = x\}$. There is a universal Turing machine U, such that for each Turing machine T there is a constant c_T, such that for all x and y, we have $K_U(x|y) \leq K_T(x|y) + c_T$. This definition is invariant up to a constant with respect to different universal Turing machines. Hence we fix a reference universal Turing machine U, and drop the subscript U by setting $K(x|y) = K_U(x|y)$. We define: The Kolmogorov complexity of a binary string x is $K(x) = K(x|\epsilon)$.

[12] A distribution is simple if it is dominated by a recursively enumerable distribution.

[13] Shallowness predicts some interesting aspects of natural language: the existence of a relatively small number of word classes, some of which are very large. See remarks on open and closed word classes below.

[14] It is currently not known what the exact size of the set of examples should be.

The relevance of shallowness for efficient language learning is clear from the following theorem:

Theorem 4 *Under* $\boldsymbol{m}$*, sentences from shallow languages can be generated/sampled in polynomial time.*

If $\mathbf{m}$ is a good approximation of co-operative linguistic behavior then the possibility that an arbitrary speaker selects a certain binary string decreases exponentially with the complexity of this string and if a language is shallow this increases the probability that the language can be learned efficiently.

Subsequent investigations using the EMILE tool has lead to the surprising insight that existing text corpora are by nature unsuitable for syntactic induction because they are semantically biased. Since nouns can be substituted for other nouns, one would expect that it would be easy to learn from a text on biology that 'whale' and 'shark' are both nouns. However, this is not the case, since the distributional evidence that they can be substituted for each other salva beneformatione is not supported by the text. Based on the semantic bias, one might learn that a whale is a mammal and that a shark is a fish, which indicates that this approach can be deployed for the induction of semantic structures, i.e., ontologies.

At the current state of research there is no good synthesis between these approaches available. Interesting problems are a better understanding of the exact nature of the language classes that are learned by existing Grammar Induction implementations and a better merge of PAC and PAC-s learnability approaches with the recursive models of the researchers working in the Gold paradigm.

3 An initial empirical model

The efficiency of human language acquisition is ill-understood at this moment. For example, there is much debate amongst linguists about the nature of universal linguistic categories and there is almost no agreement on general principles. Even the universality of the distinction between nouns and verbs has been doubted.[15] Also the value of negative results concerning the formal unlearnability of important classes of languages [Gold, 1967; Angluin, 1988; Pitt and Warmuth, 1988] as providing proof of unsurmountable barriers for natural language acquisition from scratch is very debatable. There is an early result by Horning [1969] that probabilistic variants of context-free grammars can be identified in the limit. Recent research has come up with Grammar Induction algorithms that perform reasonably well on real life corpora for linguistically interesting classes of grammars ([Seginer, 2003], EMILE: [Vervoort, 2000], Alignment-Based Learning: [van Zaanen, 2002], Evidence based State Merging: [Lang et al., 1998]). There is a large body of research on constraints, probability distributions and bias that make linguistically relevant classes of languages learnable. [Osherson et al., 1997; de la Higuera et al., 2003] A working hypothesis is that human language must be explained in terms of a structure that:

- facilitates efficient on-line interpretation,

[15] See http://linguistics.arizona.edu/~carnie/papers/V1Volume/Gil.pdf

- has adequate descriptive power,
- satisfies certain physical and biological constraints.

Using insights from algorithmic information theory, complexity theory and statistics some very general features of human languages can already be predicted (cf. the shallowness constraint proposed by Adriaans [1992, 2001]). A working hypothesis in this case is that the learning process:

- is phased,
- is hybrid, in the sense that different learning strategies might be invoked in different phases,
- is cumulative, in the sense that structures learned in phase n can be used to bootstrap structures in phase $n + 1$.[16]

Various forms of bootstrapping (prosodic, syntactic, semantic) seem to come in play in language acquisition [van Kampen, 2004]. The algorithmic aspects of these bootstrapping processes are ill-understood at the moment, although some of the Grammar Induction algorithms implicitly use them (EMILE, ABL).

An important observation in this context is the existence of open and closed word classes in natural languages. *Open classes* are N nouns, V verbs and Adj adjectives. *Closed classes* or functional categories are P prepositions, D determiners and Aux auxiliaries. Open classes can be extended without changing the character of the languages and they can be paraphrased (i.e., a cow is a mammal about the size of a horse with hooves and horns). This is not possible for closed classes (try to paraphrase the meaning of the words 'in' or 'the'). The closed classes are limited in size (about a dozen each) and changing them affects the structure of the language. Languages vary considerably in terms of functional categories. Parallels and differences between human languages are due to the differences and parallels between these functional categories. For example, in some languages the role of functional categories is partly taken over by morphology (e.g., Latin). The main problem of learning a language seems to be finding out how a language uses functional categories and by what order and speed these can be learned. Although every language seems to have an unlimited bootstrapping capability for the introduction of members of open classes, functional categories can not be learned easily in this way. A working hypothesis for first language acquisition could be that the functional categories are learned on the basis of samples taken from limited sets of open classes. Initial studies by Pinker [1999] (semantic bootstrapping) and van Kampen (syntactic bootstrapping) look promising, but more empirical research is necessary.

There is no evidence that phases of linguistic development are universal for all languages or even that subjects that learn the same language will take the same learning route. On the contrary, there is reason to suppose that, for example, in the

[16] Note that this model does not necessarily imply that linguistic performance itself is growing in a monotone way. Structures that are left unanalyzed in phase n can be overfitted (overtrained) in phase $n + 1$.

case of hearing impaired subjects some bootstrapping clues are missed because of deficiencies in the input signal. This data together with the circumstance that other communicative systems, like sign language, might have been acquired, leads to the hypothesis that these subjects follow non-standard learning routes. A better insight in these phenomena will be the basis for better diagnostic techniques for deficiencies in language development. A rough initial sketch of the various stages in children's language acquisition together with a proposal for their algorithmic structure is [Clark, 2002]:

Period	Description	Model	Learning strategy
0–9 months	Linking acoustics and events babbling	DFA	evidence-based state merging, prosodic bootstrapping
9–24 months	Children categorize words into word classes and show evidence of early sensitivity to syntax word classes	complex interaction between deixis and babbling	syntactic and semantic bootstrapping
2-3,5 years	Language meaning and syntax structure is acquired, emergence of recursive rules	context-free grammar induction as first approximation	Seginer, EMILE, ABL

Cognitive neuroscience has identified interesting ERP phenomena such as the N400 effect that is associated with semantic processing and the P600/SPS effect that has a relation with the syntactic analysis in the human brain. In recent years various neurocomputational models of syntactic processing have been proposed. Hagoort [2003] presents a neurocomputational model of syntactic processing based on a computational model of parsing due to Vosse and Kempen [2000]. These proposals for a solution of the so-called binding problem for language also have consequences of the study of human language acquisition: can the grammatical model underlying the binding solution be learned and, if so, in which way does it develop in the brain?

4 Conclusion

In this article, we gave an overview of different approaches to Computational Grammar Induction (CGI), concentrating on the field of linguistics. This includes research on the learnability of grammar classes, learning from real life linguistic data, and probabilistic approaches to Grammar Induction. We argued that researchers working on these closely related fields have been working mostly in isolation, without much interaction between the fields.

Overseeing all the developments one can not but conclude that the time for an integrative effort concerning language acquisition seems right. A cross-fertilization between linguistics, theory of computation and cognitive neuroscience might lead to breakthroughs in one or more of these fields.

References

Adriaans, P. (2001). Learning shallow context-free languages under simple distributions. In Copestake, A. and (eds.), K. V., editors, *Algebras, Diagrams and Decisions in Language, Logic and Computation*. CSLI/CUP.

Adriaans, P. and Vervoort, M. (2002). The EMILE 4.1 grammar induction toolbox. In Adriaans, P., Fernau, H., and van Zaanen, M., editors, *Grammatical Inference: Algorithms and Applications; 6th International Colloquium, ICGI 2002*, volume 2484 of *LNCS/LNAI*, pages 293–295. Springer.

Adriaans, W. P. (1992). *Language Learning from a Categorial Perspective*. PhD thesis, Universiteit van Amsterdam.

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

Angluin, D. (1982). Inference of reversible languages. *Journal of the Association for Computing Machinery*, 29(3):741–765.

Angluin, D. (1987). Learning k-bounded context-free grammars. Technical Report YALEU/DCS/TR-557, Yale University.

Angluin, D. (1988). Queries and concept learning. machine learning. *Machine Learning*, 2:319–342.

Angluin, D., Krikis, M., Sloan, R. H., and Turán, G. (1997). Malicious omissions and errors in answers to membership queries. *Machine Learning*, 28(2–3):211–255.

Charniak, E. (1993). *Statistical Language Learning*. Massachusetts Institute of Technology Press, Cambridge:MA, USA and London, UK.

Clark, E. V. (2002). *First language acquisition*. Cambridge University Press, Cambridge, UK.

de la Higuera, C., Adriaans, P., van Zaanen, M., and Oncina, J., editors (2003). *Proceedings of the Workshop and Tutorial on Learning Context-Free Grammars held at the 14th European Conference on Machine Learning (ECML) and the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD); Dubrovnik, Croatia*.

Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 10:447–474.

Hagoort, P. (2003). How the brain solves the binding problem for language: a neurocomputational model of syntactic processing. *Neuroimage*, 20(Supplement 1):S18–S29.

Horning, J. J. (1969). *A study of grammatical inference*. PhD thesis, Stanford University, Stanford:CA, USA.

Hume, D. (1909). *An Enquiry Concerning Human Understanding*, volume Vol. XXXVII, Part 3 of *The Harvard Classics*. P.F. Collier & Son.

Huybrechts, R. (1984). The weak adequacy of context-free phrase structure grammar. In de Haan, G., Trommelen, M., and Zonneveld, W., editors, *Van periferie naar kern*, pages 81–99. Foris, Dordrecht, the Netherlands.

Ishiszaka, H. (1990). Polynomial time learnability of simple deterministic languages. *Machine Learning*, 5:151.

Kanazawa, M. (1995). *Learnable classes of categorial grammars*. PhD thesis, Stanford University.

Lang, K., Pearlmutter, B., and Price, R. (1998). Results of the Abbadingo One DFA learning competition and a new evidence-driven state merging algorithm. In Honavar, V. and Slutzki, G., editors, *Grammatical Inference; 4th International Colloquium, ICGI-98*, volume 1433 of *LNCS/LNAI*, pages 1–12. Springer.

Li, M. and Vitányi, P. M. B. (1991). Learning simple concepts under simple distributions. *SIAM Journal of Computing*, 20(5):911–935.

Osherson, D., de Jongh, D., Martin, E., and Weinstein, S. (1997). *Handbook of Logic and Language*, chapter Formal Learning Theory, pages 737–775. Elsevier Science B.V.

Pinker, S. (1999). *Words and Rules: The Ingredients of Language*. Weidenfeld and Nicolson, London.

Pitt, L. and Warmuth, M. (1988). Reductions among prediction problems: On the difficulty of predicting automata. In *3rd Conference on Structure in Complexity Theory*, pages 60–69.

Sakakibara, Y. (1992). Efficient learning of context-free grammars from positive structural examples. *Information and Computation*, 97:23–60.

Seginer, Y. (2003). Learning context free grammars in the limit aided by the sample distribution. In de la Higuera et al. [2003], pages 77–88.

Shieber, S. (1985). Evidence against the context-freeness of natural language. *Linguistics and Philosophy*, 8(3):333–343.

Sokolov, J. and Snow, C. (1994). The changing role of negative evidence in theories of language development. In Gallaway, C. and Richards, B., editors, *Input and Interaction in Language Acquisition*, pages 38–55. Cambridge University Press, Cambridge, UK.

Solomonoff, R. J. (1997). The discovery of algorithmic probability. *Journal of Computer and System Sciences*, 55(1):73–88.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the Association for Computing Machinery*, 27(11):1134–1142.

van Kampen, J. (2004). Language specific bootstraps for UG categories. *International Journal of Bilingualism*. To appear.

van Zaanen, M. (2002). *Bootstrapping Structure into Language: Alignment-Based Learning*. PhD thesis, University of Leeds, Leeds, UK.

van Zaanen, M. and Adriaans, P. (2001). Alignment-Based Learning versus EMILE: A comparison. In *Proceedings of the Belgian-Dutch Conference on Artificial Intelligence (BNAIC); Amsterdam, the Netherlands*, pages 315–322.

Vervoort, M. (2000). *Games, walks and Grammars*. PhD thesis, University of Amsterdam.

Vosse, T. and Kempen, G. (2000). Syntactic structure assembly in human parsing: a computational model on competitive inhibition and lexicalist grammar. *Cognition*, 75:105–143.

Wolff, J. G. (1977). The discovery of segments in natural language. *British Journal of Psychology*, 68:97–106.

Wolff, J. G. (2003). Information compression by multiple alignment, unification and search as a unifying principle in computing and cognition. *Journal of Artificial Intelligence Research*, 19:193–230.

Yokomori, T. (2003). Polynomial-time identification of very simple grammars from positive data. *Theoretical Computer Science*, 1(298):179–206.